European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



# The sixth Visual Object Tracking VOT2018 challenge results

Matej Kristan[1], Aleš Leonardis[2], Jiří Matas[3], Michael Felsberg[4], Roman Pflugfelder[5,6], Luka Čehovin Zajc[1], Tomáš Vojíř[3], Goutam Bhat[4], Alan Lukežič[1], Abdelrahman Eldesokey[4], Gustavo Fernández[5], Álvaro García-Martín[44], Álvaro Iglesias-Arias[44], A. Aydin Alatan[28], Abel González-García[47], Alfredo Petrosino[54], Alireza Memarmoghadam[53], Andrea Vedaldi[55], Andrej Muhič[1], Anfeng He[27], Arnold Smeulders[48], Asanka G. Perera[57], Bo Li[7], Boyu Chen[13], Changick Kim[24], Changsheng Xu[30], Changzhen Xiong[9], Cheng Tian[16], Chong Luo[27], Chong Sun[13], Cong Hao[52], Daijin Kim[34], Deepak Mishra[19], Deming Chen[52], Dong Wang[13], Dongyoon Wee[31], Efstratios Gavves[48], Erhan Gundogdu[14], Erik Velasco-Salido[44], Fahad Shahbaz Khan[4], Fan Yang[42], Fei Zhao[32,50], Feng Li[16], Francesco Battistone[26], George De Ath[51], Gorthi R. K. S. Subrahmanyam[19], Guilherme Bastos[45], Haibin Ling[42], Hamed Kiani Galoogahi[35], Hankyeol Lee[24], Haojie Li[40], Haojie Zhao[13], Heng Fan[42], Honggang Zhang[10], Horst Possegger[15], Houqiang Li[56], Huchuan Lu[13], Hui Zhi[9], Huiyun Li[39], Hyemin Lee[34], Hyung Jin Chang[2], Isabela Drummond[45], Jack Valmadre[55], Jaime Spencer Martin[58], Javaan Chahl[57], Jin Young Choi[37], Jing Li[12], Jinqiao Wang[32,50], Jinqing Qi[13], Jinyoung Sung[31], Joakim Johnander[4], Joao Henriques[55], Jongwon Choi[37], Joost van de Weijer[47], Jorge Rodríguez Herranz[1,41], José M. Martínez[44], Josef Kittler[58], Junfei Zhuang[8,10], Junyu Gao[30], Klemen Grm[1], Lichao Zhang[47], Lijun Wang[13], Lingxiao Yang[17], Litu Rout[19], Liu Si[22], Luca Bertinetto[55], Lutao Chu[39,50], Manqiang Che[9], Mario Edoardo Maresca[54], Martin Danelljan[4], Ming-Hsuan Yang[49], Mohamed Abdelpakey[25], Mohamed Shehata[25], Myunggu Kang[31], Namhoon Lee[55], Ning Wang[56], Ondrej Miksik[55], P. Moallem[53], Pablo Vicente-Moñivar[44], Pedro Senna[46], Peixia Li[13], Philip Torr[55], Priya Mariam Raju[19], Qian Ruihe[22], Qiang Wang[30], Qin Zhou[38], Qing Guo[43], Rafael Martín-Nieto[44], Rama Krishna Gorthi[19], Ran Tao[48], Richard Bowden[58], Richard Everson[51], Runling Wang[33], Sangdoo Yun[37], Seokeon Choi[24], Sergio Vivas[44], Shuai Bai[8,10], Shuangping Huang[40], Sihang Wu[40], Simon Hadfield[58], Siwen Wang[13], Stuart Golodetz[55], Tang Ming[32,50], Tianyang Xu[23], Tianzhu Zhang[30], Tobias Fischer[18], Vincenzo Santopietro[54], Vitomir Štruc[1], Wang Wei[11], Wangmeng Zuo[16], Wei Feng[43], Wei Wu[36], Wei Zou[21], Weiming Hu[30], Wengang Zhou[56], Wenjun Zeng[27], Xiaofan Zhang[52], Xiaohe Wu[16], Xiao-Jun Wu[23], Xinmei Tian[56], Yan Li[9], Yan Lu[9], Yee Wei Law[57], Yi Wu[20,29], Yiannis Demiris[18], Yicai Yang[40], Yifan Jiao[30], Yuhong Li[10,52], Yunhua Zhang[13], Yuxuan Sun[13], Zheng Zhang[59], Zheng Zhu[21,50], Zhen-Hua Feng[58], Zhihui Wang[13], and Zhiqun He[8,10]

[1] University of Ljubljana, Slovenia
[2] University of Birmingham, United Kingdom
[3] Czech Technical University, Czech Republic
[4] Linköping University, Sweden

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



[5] Austrian Institute of Technology, Austria
[6] TU Wien, Austria
[7] Beihang University, China
[8] Beijing Faceall Co., China
[9] Beijing Key Laboratory of Urban Intelligent Control, China
[10] Beijing University of Posts and Telecommunications, China
[11] China Huayin Ordnance Test Center, China
[12] Civil Aviation University Of China, China
[13] Dalian University of Technology, China
[14] EPFL, Switzerland
[15] Graz University of Technology, Austria
[16] Harbin Institute of Technology, China
[17] Hong Kong Polytechnic University, Hong Kong
[18] Imperial College London, United Kingdom
[19] Indian Institute of Space Science and Technology, India
[20] Indiana University, USA
[21] Institute of Automation, Chinese Academy of Sciences, China
[22] Institute of Information Engineering, China
[23] Jiangnan University, China
[24] KAIST, South Korea
[25] Memorial University of Newfoundland, Canada
[26] Mer Mec S.p.A., Italy
[27] Microsoft Research Asia, China
[28] Middle East Technical University, Turkey
[29] Nanjing Audit University, China
[30] National Laboratory of Pattern Recognition, China
[31] Naver Corporation, South Korea
[32] NLPR, Institute of Automation, Chinese Academy of Sciences, China
[33] North China University of Technology, China
[34] POSTECH, South Korea
[35] Robotics Institute, Carnegie Mellon University, USA
[36] Sensetime, China
[37] Seoul National University, South Korea
[38] Shanghai Jiao Tong University, China
[39] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[40] South China University of Technology, China
[41] Technical University of Madrid, Spain
[42] Temple University, USA
[43] Tianjin University, China
[44] Universidad Autónoma de Madrid, Spain
[45] Universidade Federal de Itajubá, Brazil
[46] Universidade Federal do Mato Grosso do Sul, Brazil
[47] Universitat Autónoma de Barcelona, Spain
[48] University of Amsterdam, Netherlands
[49] University of California, USA
[50] University of Chinese Academy of Sciences, China
[51] University of Exter, United Kingdom
[52] University of Illinois Urbana-Champaign, USA
[53] University of Isfahan, Iran
[54] University of Naples Parthenope, Italy

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1

[55] University of Oxford, United Kingdom
[56] University of Science and Technology of China, China
[57] University of South Australia, Australia
[58] University of Surrey, United Kingdom
[59] Zhejiang University, China

**Abstract.** The Visual Object Tracking challenge VOT2018 is the sixth annual tracker benchmarking activity organized by the VOT initiative. Results of over eighty trackers are presented; many are state-of-the-art trackers published at major computer vision conferences or in journals in the recent years. The evaluation included the standard VOT and other popular methodologies for short-term tracking analysis and a "real-time" experiment simulating a situation where a tracker processes images as if provided by a continuously running sensor. A long-term tracking subchallenge has been introduced to the set of standard VOT sub-challenges. The new subchallenge focuses on long-term tracking properties, namely coping with target disappearance and reappearance. A new dataset has been compiled and a performance evaluation methodology that focuses on long-term tracking capabilities has been adopted. The VOT toolkit has been updated to support both standard short-term and the new long-term tracking subchallenges. Performance of the tested trackers typically by far exceeds standard baselines. The source code for most of the trackers is publicly available from the VOT page. The dataset, the evaluation kit and the results are publicly available at the challenge website[60].

## 1 Introduction

Visual object tracking has consistently been a popular research area over the last two decades. The popularity has been propelled by significant research challenges tracking offers as well as the industrial potential of tracking-based applications. Several initiatives have been established to promote tracking, such as PETS [95], CAVIAR[61], i-LIDS [62], ETISEO[63], CDC [25], CVBASE [64], FERET [67], LTDT [65], MOTC [44,76] and Videonet [66], and since 2013 short-term single target visual object tracking has been receiving a strong push toward performance evaluation standardisation from the VOT [60] initiative. The primary goal of VOT is establishing datasets, evaluation measures and toolkits as well as creating a platform for discussing evaluation-related issues through organization of tracking challenges. Since 2013, five challenges have taken place in conjunction with

[60] http://votchallenge.net
[61] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[62] http://www.homeoffice.gov.uk/science-research/hosdb/i-lids
[63] http://www-sop.inria.fr/orion/ETISEO
[64] http://vision.fe.uni-lj.si/cvbase06/
[65] http://www.micc.unifi.it/LTDT2014/
[66] http://videonet.team

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



ICCV2013 (VOT2013 [41]), ECCV2014 (VOT2014 [42]), ICCV2015 (VOT2015 [40]), ECCV2016 (VOT2016 [39]) and ICCV2017 (VOT2017 [38]).

This paper presents the VOT2018 challenge, organized in conjunction with the ECCV2018 Visual Object Tracking Workshop, and the results obtained. The VOT2018 challenge addresses two classes of trackers. The first class has been considered in the past five challenges: single-camera, single-target, model-free, causal trackers, applied to short-term tracking. The *model-free* property means that the only training information provided is the bounding box in the first frame. The *short-term* tracking means that trackers are assumed not to be capable of performing successful re-detection after the target is lost and they are therefore reset after such an event. *Causality* requires that the tracker does not use any future frames, or frames prior to re-initialization, to infer the object position in the current frame. The second class of trackers is introduced this year in the *first VOT long-term sub-challenge*. This subchallenge considers single-camera, single-target, model-free long-term trackers. The *long-term* tracking means that the trackers are *required* to perform re-detection after the target has been lost and are therefore *not* reset after such an event. In the following, we overview the most closely related works and point out the contributions of VOT2018.

### 1.1 Related work in short-term tracking

A lot of research has been invested into benchmarking and performance evaluation in short-term visual object tracking [41,42,40,39,38,43,83,75,92,47,51,61,96,62,101]. The currently most widely-used methodologies have been popularized by two benchmark papers: "Online Tracking Benchmark" (OTB) [92] and "Visual Object Tracking challenge" (VOT) [41]. The methodologies differ in the evaluation protocols as well as the performance measures.

The OTB-based evaluation approaches initialize the tracker in the first frame and let it runs until the end of the sequence. The benefit of this protocol is its implementation simplicity. But target predictions become irrelevant for tracking accuracy of short-term trackers after the initial failure, which introduces variance and bias in the results [43]. The VOT evaluation approach addresses this issue by resetting the tracker after each failure.

All recent performance evaluation protocols measure tracking accuracy primarily by intersection over union (IoU) between the ground truth and tracker prediction bounding boxes. A legacy center-based measure initially promoted by Babenko et al. [3] and later adopted by [90] is still often used, but is theoretically brittle and inferior to the overlap-based measure [83]. In the no-reset-based protocols the overall performance is summarized by the average IoU over the dataset (i.e., average overlap) [90,83]. In the VOT reset-based protocols, two measures are used to probe the performance: (i) accuracy and (ii) robustness. They measure the overlap during successful tracking periods and the number of times the tracker fails. Since 2015, the VOT primary measure is the expected average overlap (EAO) – a principled combination of accuracy and robustness. The

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



VOT reports the so-called state-of-the-art bound (*SotA* bound) on all their annual challenges. Any tracker exceeding *SotA* bound is considered state-of-the-art by VOT standard. This bound was introduced to counter the trend of considering state-of-the-art only those trackers that rank number one on benchmarks. By *SotA* bound, the hope was to remove the need of fine-tuning to benchmarks and to incent community-wide exploration of a wider spectrum of trackers, not necessarily getting the number one rank.

Tracking speed was recognized as an important tracking factor in VOT2014 [42]. Initially the speed was measured in terms of equivalent filtering operations [42] to reduce the varying hardware influence. This measure was abandoned due to limited normalization capability and due to the fact that speed often varies a lot during tracking. Since VOT2017 [42] speed aspects are measured by a protocol that requires real-time processing of incoming frames.

Most tracking datasets [92,47,75,51,61] have partially followed the trend in computer vision of increasing the number of sequences. But quantity does not necessarily reflect diversity nor richness in attributes. Over the years, the VOT [41,42,40,43,39,38] has developed a dataset construction methodology for constructing moderately large challenging datasets from a large pool of sequences. Through annual discussions at VOT workshops, the community expressed a request for evaluating trackers on a sequestered dataset. In response, the VOT2017 challenge introduced a sequestered dataset evaluation for winner identification in the main short-term challenge. In 2015 VOT introduced a sub-challenge for evaluating short-term trackers on thermal and infra-red sequences (VOT-TIR2015) with a dataset specially designed for that purpose [21]. Recently, datasets focusing on various short-term tracking aspects have been introduced. The UAV123 [61] and [101] proposed datasets for tracking from drones. Lin et al. [94] proposed a dataset for tracking faces by mobile phones. Galoogahi et al. [22] introduced a high-frame-rate dataset to analyze trade-offs between tracker speed and robustness. Čehovin et al. [96] proposed a dataset with an active camera view control using omni directional videos. Mueller et al. [62] recently re-annotated selected sequences from Youtube bounding boxes [69] to consider tracking in the wild. Despite significant activity in dataset construction, the VOT dataset remains unique for its carefully chosen and curated sequences guaranteeing relatively unbiased assessment of performance with respect to attributes.

### 1.2 Related work in long-term tracking

Long-term (LT) trackers have received far less attention than short-term (ST) trackers. A major difference between ST and LT trackers is that LT trackers are required to handle situations in which the target may leave the field of view for a longer duration. This means that LT trackers have to detect target absence and re-detect the target when it reappears. Therefore a natural evaluation protocol for LT tracking is a no-reset protocol.

A typical structure of a long-term tracker is a short-term component with a relatively small search range responsible for frame-to-frame association and a

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



detector component responsible for detecting target reappearance. In addition, an interaction mechanism between the short-term component and the detector is required that appropriately updates the visual models and switches between target tracking and detection. This structure originates from two seminal papers in long-term tracking TLD [37] and Alien [66], and has been reused in all subsequent LT trackers (e.g., [59,65,34,100,57,20]).

The set of performance measures in long-term tracking is quite diverse and has not been converging like in the short-term tracking. The early long-term tracking papers [37,66] considered measures from object detection literature since detectors play a central role in LT tracking. The primary performance measures were precision, recall and F-measure computed at 0.5 IoU (overlap) threshold. But for tracking, the overlap of 0.5 is over-restrictive as discussed in [37,43] and does not faithfully reflect the overall tracking capabilities. Furthermore, the approach requires a binary output – either target is present or absent. In general, a tracker can report the target position along with a presence certainty score which offers a more accurate analysis, but this is prevented by the binary output requirement. In addition to precision/recall measures, the authors of [37,66] proposed using average center error to analyze tracking accuracy. But center-error-based measures are even more brittle than IoU-based measures, are resolution-dependent and are computed only in frames where the target is present and the tracker reports its position. Thus most papers published in the last few years (e.g, [34,57,20]) have simply used the short-term average overlap performance measure from [90,61]. But this measure does not account for the tracker's ability to correctly report target absence and favors reporting target positions at every frame. Attempts were made to address this drawback [79,60] by specifying an overlap equal to 1 when the tracker correctly predicts the target absence, but this does not clearly separate re-detection ability from tracking accuracy. Recently, Lukežič et. al. [56] have proposed *tracking* precision, *tracking* recall and *tracking* F-measure that avoid dependence on the IoU threshold and allow analyzing trackers with presence certainty outputs without assuming a predefined scale of the outputs. They have shown that their primary measure, the tracking F-measure, reduces to a standard short-term measure (average overlap) when computed in a short-term setup.

Only few datasets have been proposed in long-term tracking. The first dataset was introduced by the LTDT challenge [65], which offered a collection of specific videos from [37,66,45,75]. These videos were chosen using the following definition of the long-term sequence: *"long-term sequence is a video that is at least 2 minutes long (at 25-30 fps), but ideally 10 minutes or longer"* [65]. Mueller et al. [61] proposed a UAV20L dataset containing twenty long sequences with many target disappearances recorded from drones. Recently, three benchmarks that propose datasets with many target disappearances have almost concurrently appeared on pre-pub [60,56,36]. The benchmark [60] primarily analyzes performance of short-term trackers on long sequences, and [36] proposes a huge dataset constructed from Youtube bounding boxes [69]. To cope with significant dataset size, [36] annotate the tracked object every few frames. The benchmark [60] does not dis-

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



tinguish between short-term and long-term trackers architectures but considers LT tracking as the ability to track long sequences attributing most of performance boosts to robust visual models. The benchmarks [36,56], on the other hand, point out the importance of re-detection and [56] uses this as a guideline to construct a moderately sized dataset with many long-term specific attributes. In fact, [56] argue that long-term tracking does not just refer to the sequence length, but more importantly to the sequence properties (number of target disappearances, etc.) and the type of tracking output expected. They argue that there are several levels of tracker types between pure short-term and long-term trackers and propose a new short-term/long-term tracking taxonomy covering four classes of ST/LT trackers. For these reasons, we base the VOT long-term dataset and evaluation protocols described in Section 3 on [56].

### 1.3 The VOT2018 challenge

VOT2018 considers short-term as well as long-term trackers in separate subchallenges. The evaluation toolkit and the datasets are provided by the VOT2018 organizers. These were released on April 26th 2018 for beta-testing. The challenge officially opened on May 5th 2018 with approximately a month available for results submission.

The authors participating in the challenge were required to integrate their tracker into the VOT2018 evaluation kit, which automatically performed a set of standardized experiments. The results were analyzed according to the VOT2018 evaluation methodology.

Participants were encouraged to submit their own new or previously published trackers as well as modified versions of third-party trackers. In the latter case, modifications had to be significant enough for acceptance. Participants were expected to submit a single set of results per tracker. Changes in the parameters did not constitute a different tracker. The tracker was required to run with fixed parameters in all experiments. The tracking method itself was allowed to internally change specific parameters, but these had to be set automatically by the tracker, e.g., from the image size and the initial size of the bounding box, and were not to be set by detecting a specific test sequence and then selecting the parameters that were hand-tuned for this sequence.

Each submission was accompanied by a short abstract describing the tracker, which was used for the short tracker descriptions in Appendix A. In addition, participants filled out a questionnaire on the VOT submission page to categorize their tracker along various design properties. Authors had to agree to help the VOT technical committee to reproduce their results in case their tracker was selected for further validation. Participants with sufficiently well-performing submissions, who contributed with the text for this paper and agreed to make their tracker code publicly available from the VOT page were offered co-authorship of this results paper.

To counter attempts of intentionally reporting large bounding boxes to avoid resets, the VOT committee analyzed the submitted tracker outputs. The commit-

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



tee reserved the right to disqualify the tracker should such or a similar strategy be detected.

To compete for the winner of VOT2018 challenge, learning from the tracking datasets (OTB, VOT, ALOV, NUSPRO and TempleColor) was prohibited. The use of class labels specific to VOT was not allowed (i.e., identifying a target class in each sequence and applying pre-trained class-specific trackers is not allowed). An agreement to publish the code online on VOT webpage was required. The organizers of VOT2018 were allowed to participate in the challenge, but did not compete for the winner of the VOT2018 challenge title. Further details are available from the challenge homepage[67].

Like VOT2017, the VOT2018 was running the main VOT2018 short-term sub-challenge and the VOT2018 short-term real-time sub-challenge, but did not run the short-term thermal and infrared VOT-TIR sub-challenge. As a significant novelty, the VOT2018 introduces a new VOT2018 long-term tracking challenge, adopting the methodology from [56]. The VOT2018 toolkit has been updated to allow seamless use in short-term and long-term tracking evaluation. In the following we overview the sub-challenges.

## 2 The VOT2018 short-term challenge

The VOT2018 short-term challenge contains the main VOT2018 short-term sub-challenge and the VOT2018 realtime sub-challenge. Both sub-challenges used the same dataset, but different evaluation protocols.

The VOT2017 results have indicated that the 2017 dataset has not saturated, therefore *the dataset was used unchanged* in the VOT2018 short-term challenge. The dataset contains 60 sequences released to public (i.e., VOT2017 *public* dataset) and another 60 *sequestered* sequences (i.e., VOT2017 *sequestered* dataset). Only the former dataset was released to the public, while the latter was not disclosed and was used only to identify the winner of the main VOT2018 short-term challenge. The target in the sequences is annotated by a rotated bounding box and all sequences are per-frame annotated by the following visual attributes: (i) occlusion, (ii) illumination change, (iii) motion change, (iv) size change and (v) camera motion. Frames that did not correspond to any of the five attributes were denoted as (vi) unassigned.

### 2.1 Performance measures and evaluation protocol

As in VOT2017 [38], three primary measures were used to analyze the short-term tracking performance: accuracy ($A$), robustness ($R$) and expected average overlap (EAO). In the following, these are briefly overviewed and we refer to [40,43,83] for further details.

The VOT short-term challenges apply a reset-based methodology. Whenever a tracker predicts a bounding box with zero overlap with the ground truth, a

[67] http://www.votchallenge.net/vot2018/participation.html

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



failure is detected and the tracker is re-initialized five frames after the failure. Accuracy and robustness [83] are the basic measures used to probe tracker performance in the reset-based experiments. The accuracy is the average overlap between the predicted and ground truth bounding boxes during successful tracking periods. The robustness measures how many times the tracker loses the target (fails) during tracking. The potential bias due to resets is reduced by ignoring ten frames after re-initialization in the accuracy measure (note that a tracker is reinitialized five frames after failure), which is quite a conservative margin [43]. Average accuracy and failure-rates are reported for stochastic trackers, which are run 15 times.

The third, primary measure, called the expected average overlap (EAO), is an estimator of the average overlap a tracker is expected to attain on a large collection of short-term sequences with the same visual properties as the given dataset. The measure addresses the problem of increased variance and bias of AO [92] measure due to variable sequence lengths. Please see [40] for further details on the average expected overlap measure. For reference, the toolkit also ran a no-reset experiment and the AO [92] was computed (available in the online results).

### 2.2 The VOT2018 real-time sub-challenge

The VOT2018 real-time sub-challenge was introduced in VOT2017 [38] and is a variation of the main VOT2018 short-term sub-challenge. The main VOT2018 short-term sub-challenge does not place any constraint on the time for processing a single frame. In contrast, the VOT2018 real-time sub-challenge requires predicting bounding boxes faster or equal to the video frame-rate. The toolkit sends images to the tracker via the Trax protocol [10] at 20fps. If the tracker does not respond in time, the last reported bounding box is assumed as the reported tracker output at the available frame (zero-order hold dynamic model).

The toolkit applies a reset-based VOT evaluation protocol by resetting the tracker whenever the tracker bounding box does not overlap with the ground truth. The VOT frame skipping is applied as well to reduce the correlation between resets.

### 2.3 Winner identification protocol

On the main VOT2018 short-term sub-challenge, the winner is identified as follows. Trackers are ranked according to the EAO measure on the public dataset. Top ten trackers are re-run by the VOT2018 committee on the sequestered dataset. The top ranked tracker on the sequestered dataset not submitted by the VOT2018 committee members is the winner of the main VOT2018 short-term challenge. The winner of the VOT2018 real-time challenge is identified as the top-ranked tracker not submitted by the VOT2018 committee members according to the EAO on the public dataset.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



## 3 The VOT2018 long-term challenge

The VOT2018 long-term challenge focuses on the long-term tracking properties. In a long-term setup, the object may leave the field of view or become fully occluded for a long period. Thus in principle, a tracker is required to report the target absence. To make the integration with the toolkit compatible with the short-term setup, we require the tracker to report the target position in each frame and provide a confidence score of target presence. The VOT2018 adapts long-term tracker definitions, dataset and the evaluation protocol from [56]. We summarize these in the following and direct the reader to the original paper for more details.

### 3.1 The short-term/long-term tracking spectrum

The following definitions from [56] are used to position the trackers on the short-term/long-term spectrum:

1. **Short-term tracker** ($\mathrm{ST}_0$). The target position is reported at each frame. The tracker does not implement target re-detection and does not explicitly detect occlusion. Such trackers are likely to fail at the first occlusion as their representation is affected by any occluder.
2. **Short-term tracker with conservative updating** ($\mathrm{ST}_1$). The target position is reported at each frame. Target re-detection is not implemented, but tracking robustness is increased by selectively updating the visual model depending on a tracking confidence estimation mechanism.
3. **Pseudo long-term tracker** ($\mathrm{LT}_0$). The target position is not reported in frames when the target is not visible. The tracker does not implement explicit target re-detection but uses an internal mechanism to identify and report tracking failure.
4. **Re-detecting long-term tracker** ($\mathrm{LT}_1$). The target position is not reported in frames when the target is not visible. The tracker detects tracking failure and implements explicit target re-detection.

### 3.2 The dataset

Trackers are evaluated on the LTB35 dataset [56]. This dataset contains 35 sequences, carefully selected to obtain a dataset with long sequences containing many target disappearances. Twenty sequences were obtained from the UAVL20 [61], three from [37], six sequences were taken from Youtube and six sequences were generated from the omnidirectional view generator AMP [96] to ensure many target disappearances. Sequence resolutions range between $1280\times720$ and $290\times217$. The dataset contains 14687 frames, with 433 target disappearances. Each sequence contains on average 12 long-term target disappearances, each lasting on average 40 frames.

The targets are annotated by axis-aligned bounding boxes. Sequences are annotated by the following visual attributes: (i) Full occlusion, (ii) Out-of-view,

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



(iii) Partial occlusion, (iv) Camera motion, (v) Fast motion, (vi) Scale change, (vii) Aspect ratio change, (viii) Viewpoint change, (ix) Similar objects. Note this is per-sequence, not per-frame annotation and a sequence can be annotated by several attributes.

### 3.3 Performance measures

We use three long-term tracking performance measures proposed in [56]: tracking precision ($Pr$), tracking recall ($Re$) and tracking F-score. These are briefly described in the following.

Let $G_t$ be the ground truth target pose, let $A_t(\tau_\theta)$ be the pose predicted by the tracker, $\theta_t$ the prediction certainty score at time-step $t$, $\tau_\theta$ be a classification (detection) threshold. If the target is absent, the ground truth is an empty set, i.e., $G_t = \emptyset$. Similarly, if the tracker did not predict the target or the prediction certainty score is below a classification threshold i.e., $\theta_t < \tau_\theta$, the output is $A_t(\tau_\theta) = \emptyset$. Let $\Omega(A_t(\tau_\theta), G_t)$ be the intersection over union between the tracker prediction and the ground truth and let $N_g$ be the number of frames with $G_t \neq \emptyset$ and $N_p$ the number of frames with existing prediction, i.e., $A_t(\tau_\theta) \neq \emptyset$.

In detection literature, the prediction matches the ground truth if the overlap $\Omega(A_t(\tau_\theta), G_t)$ exceeds a threshold $\tau_\Omega$, which makes precision and recall dependent on the minimal classification certainty as well as minimal overlap thresholds. This problem is addressed in [56] by integrating the precision and recall over all possible overlap thresholds[68]. The tracking precision and tracking recall at classification threshold $\tau_\theta$ are defined as

$$Pr(\tau_\theta) = \frac{1}{N_p} \sum_{t \in \{t : A_t(\theta_t) \neq \emptyset\}} \Omega(A_t(\theta_t), G_t), \tag{1}$$

$$Re(\tau_\theta) = \frac{1}{N_g} \sum_{t \in \{t : G_t \neq \emptyset\}} \Omega(A_t(\theta_t), G_t). \tag{2}$$

Precision and accuracy are combined into a single score by computing the tracking F-measure:

$$F(\tau_\theta) = 2Pr(\tau_\theta)Re(\tau_\theta)/(Pr(\tau_\theta) + Re(\tau_\theta)). \tag{3}$$

Long-term tracking performance can thus be visualized by tracking precision, tracking accuracy and tracking F-measure plots by computing these scores for all thresholds $\tau_\theta$.

The primary long-term tracking measure [56] is F-score, defined as the highest score on the F-measure plot, i.e., taken at the tracker-specific optimal threshold. This avoids arbitrary manual-set thresholds in the primary performance measure.

[68] Note that this can be thought of as computing the area under the curve score [90] of a precision plot computed at certainty threshold $\tau_\theta$.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### 3.4 Re-detection experiment

We also adapt an experiment from [56] designed to test the tracker's re-detection capability separately from the short-term component. This experiment generates an artificial sequence in which the target does not change appearance but only location. An initial frame of a sequence is padded with zeros to the right and down to the three times original size. This frame is repeated for the first five frames in the artificial sequence. For the remainder of the frames, the target is cropped from the initial image and placed in the bottom right corner of the frame with all other pixels set to zero.

A tracker is initialized in the first frame and the experiment measures the number of frames required to re-detect the target after position change. This experiment is re-run over artificial sequences generated from all sequences in the LTB35 dataset.

### 3.5 Evaluation protocol

A tracker is evaluated on a dataset of several sequences by initializing on the first frame of a sequence and run until the end of the sequence without re-sets. The precision-recall graph from (1) is calculated on each sequence and averaged into a single plot. This guarantees that the result is not dominated by extremely long sequences. The F-measure plot is computed according to (3) from the average precision-recall plot. The maximal score on the F-measure plot (F-score) is taken as the long-term tracking primary performance measure.

### 3.6 Winner identification protocol

The winner of the VOT2018 long-term tracking challenge is identified as the top-ranked tracker not submitted by the VOT2018 committee members according to the F-score on the LTB35 dataset.

## 4 The VOT2018 short-term challenge results

This section summarizes the trackers submitted to the VOT short-term (VOT2018 ST) challenge, results analysis and winner identification.

### 4.1 Trackers submitted

In all, 56 valid entries were submitted to the VOT2018 short-term challenge. Each submission included the binaries or source code that allowed verification of the results if required. The VOT2018 committee and associates additionally contributed 16 baseline trackers. For these, the default parameters were selected, or, when not available, were set to reasonable values. Thus in total 72 trackers were tested on the VOT2018 short-term challenge. In the following we briefly

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



overview the entries and provide the references to original papers in the Appendix A where available.

Of all participating trackers, 51 trackers (71%) were categorized as $ST_0$, 18 trackers (25%) as $ST_1$, and three (4%) as $LT_1$. 76% applied discriminative and 24% applied generative models. Most trackers – 75% – used holistic model, while 25% of the participating trackers used part-based models. Most trackers applied either a locally uniform dynamic model[69] (76%), a nearly-constant-velocity (7%), or a random walk dynamic model (15%), while only a single tracker applied a higher order dynamic model (1%).

The trackers were based on various tracking principles: 4 trackers (6%) were based on CNN matching (ALAL A.2, C3DT A.72, LSART A.40, RAnet A.57), one tracker was based on recurrent neural network (ALAL A.2), 14 trackers (18%) applied Siamese networks (ALAL A.2, DensSiam A.23, DSiam A.30, LWDNTm A.41, LWDNTthi A.42, MBSiam A.48, SA_Siam_P A.59, SA_Siam_R A.60, SiamFC A.34, SiamRPN A.35, SiamVGG A.63, STST A.66, UpdateNet A.1), 3 trackers (4%) applied support vector machines (BST A.6, MEEM A.47, struck2011 A.68), 38 trackers (53%) applied discriminative correlation filters (ANT A.3, BoVW_CFT A.4, CCOT A.11, CFCF A.13, CFTR A.15, CPT A.7, CPT_fast A.8, CSRDCF A.24, CSRTPP A.25, CSTEM A.9, DCFCF A.22, DCFNet A.18, DeepCSRDCF A.17, DeepSTRCF A.20, DFPReco A.29, DLSTpp A.28, DPT A.21, DRT A.16, DSST A.26, ECO A.31, HMMTxD A.53, KCF A.38, KFebT A.37, LADCF A.39, MCCT A.50, MFT A.51, MRSNCC A.49, R_MCPF A.56, RCO A.12, RSECF A.14, SAPKLTF A.62, SRCT A.58, SRDCF A.64, srdcf_deep A.19, srdcf_dif A.32, Staple A.67, STBACF A.65, TRACA A.69, UPDT A.71), 6 trackers (8%) applied mean shift (ASMS A.61, CPOINT A.10, HMMTxD A.53, KFebT A.37, MRSNCC A.49, SAPKLTF A.62) and 8 trackers (11%) applied optical flow (ANT A.3, CPOINT A.10, FoT A.33, Fragtrac A.55, HMMTxD A.53, LGT A.43, MRSNCC A.49, SAPKLTF A.62).

Many trackers used combinations of several features. CNN features were used in 62% of trackers – these were either trained for discrimination (32 trackers) or localization (13 trackers). Hand-crafted features were used in 44% of trackers, keypoints in 14% of trackers, color histograms in 19% and grayscale features were used in 24% of trackers.

### 4.2 The main VOT2018 short-term sub-challenge results

The results are summarized in the AR-raw plots and EAO curves in Figure 1 and the expected average overlap plots in Figure 2. The values are also reported in Table 2. The top ten trackers according to the primary EAO measure (Figure 2) are LADCF A.39, MFT A.51, SiamRPN A.35, UPDT A.71, RCO A.12, DRT A.16, DeepSTRCF A.20, SA_Siam_R A.60, CPT A.7 and DLSTpp A.28. All these trackers apply a discriminatively trained correlation filter on top of

[69] The target was sought in a window centered at its estimated position in the previous frame. This is the simplest dynamic model that assumes all positions within a search region contain the target have equal prior probability.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



multidimensional features except from SiamRPN and SA_Siam_R, which apply siamese networks. Common networks used by the top ten trackers are Alexnet, Vgg and Resnet in addition to localization pre-trained networks. Many trackers combine the deep features with HOG, Colornames and a grayscale patch.

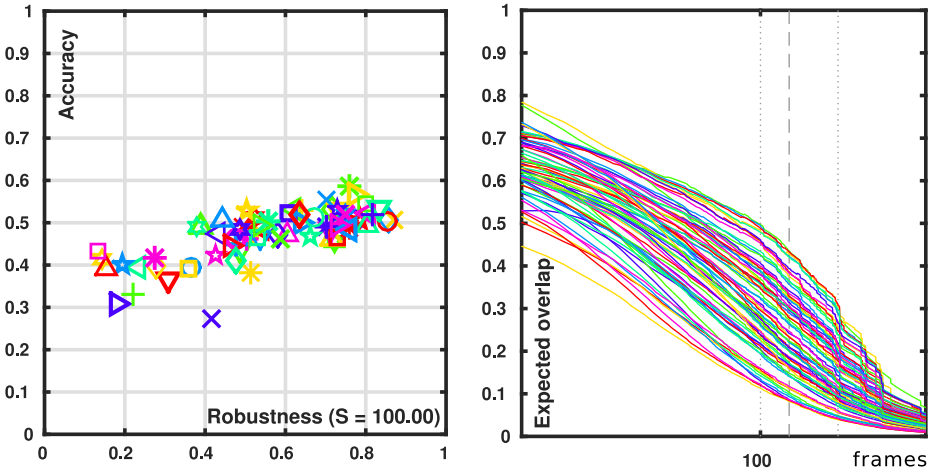


Fig. 1: The AR-raw plots generated by sequence pooling (left) and EAO curves (right).

The top performer on public dataset is LADCF (A.39). This tracker trains a low-dimensional DCF by using an adaptive spatial regularizer. Adaptive spatial regularization and temporal consistency are combined into a single objective function. The tracker uses HOG, Colournames and ResNet-50 features. Data augmentation by flipping, rotating and blurring is applied to the Resnet features. The second-best ranked tracker is MFT (A.51). This tracker adopts CFWCR [31] as a baseline feature learning algorithm and applies a continuous convolution operator [15] to fuse multiresolution features. The different resolutions are trained independently for target position prediction, which, according to the authors, significantly boosts the robustness. The tracker uses ResNet-50, SE-ResNet-50, HOG and Colornames.

The top trackers in EAO are also among the most robust trackers, which means that they are able to track longer without failing. The top trackers in robustness (Figure 1) are MFT A.51, LADCF A.39, RCO A.12, UPDT A.71, DRT A.16, LSART A.40, DeepSTRCF A.20, DLSTpp A.28, CPT A.7 and SA_Siam_R A.60. On the other hand, the top performers in accuracy are SiamRPN A.35, SA_Siam_R A.60, FSAN A.70, DLSTpp A.28, UPDT A.71, MCCT A.50, SiamVGG A.63, ALAL A.2, DeepSTRCF A.20 and SA_Siam_P A.59.

The trackers which have been considered as baselines or state-of-the-art even few years ago, i.e., MIL (A.52), IVT (A.36), Struck [28] and KCF (A.38) are positioned at the lower part of the AR-plots and at the tail of the EAO rank list. This speaks of the significant quality of the trackers submitted to VOT2018. In fact, 19 tested trackers (26%) have been recently (2017/2018) published at com-

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



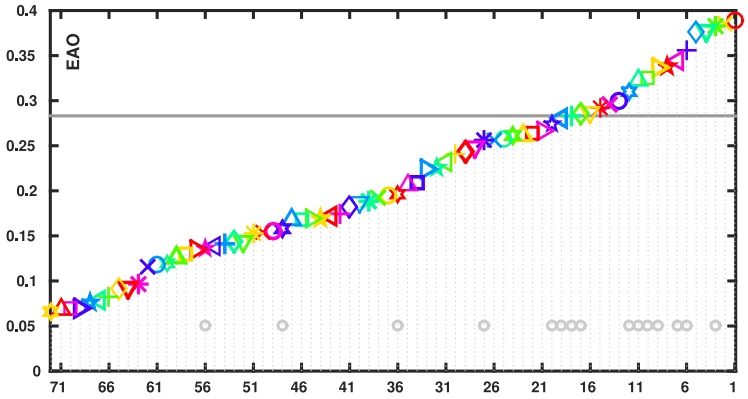


Fig. 2: Expected average overlap graph with trackers ranked from right to left. The right-most tracker is the top-performing according to the VOT2018 expected average overlap values. The dashed horizontal line denotes the average performance of ten state-of-the-art trackers published in 2017 and 2018 at major computer vision venues. These trackers are denoted by gray circle in the bottom part of the graph.

puter vision conferences and journals. These trackers are indicated in Figure 2, along with their average performance, which constitutes a very strict VOT2018 state-of-the-art bound. Approximately 26% of submitted trackers exceed this bound.

| | **CM** | **IC** | **MC** | **OC** | **SC** |
|---|---|---|---|---|---|
| Accuracy | 0.49 | 0.47 | 0.47 ③ | 0.40 ① | 0.43 ② |
| Robustness | 0.74 | 1.05 ② | 0.87 ③ | 1.19 ① | 0.61 |

Table 1: Tracking difficulty with respect to the following visual attributes: camera motion (CM), illumination change (IC), motion change (MC), occlusion (OC) and size change (SC).

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



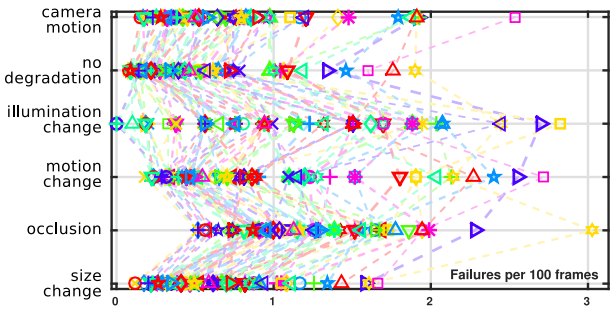


Fig. 3: Failure rate with respect to the visual attributes.

| | | Baseline | | | Realtime | | | Unsup. | |
|---|---|---|---|---|---|---|---|---|---|
| | **Tracker** | **EAO** | **A** | **R** | **EAO** | **A** | **R** | **AO** | **Impl.** |
| 1. | LADCF | 0.389 ① | 0.503 | 0.159 ③ | 0.066 | 0.314 | 1.358 | 0.421 | D M C |
| 2. | MFT | 0.385 ② | 0.505 | 0.140 ① | 0.060 | 0.337 | 1.592 | 0.393 | D M G |
| 3. | SiamRPN | 0.383 ③ | 0.586 ① | 0.276 | 0.383 ① | 0.586 ① | 0.276 ② | 0.472 ② | D P G |
| 4. | UPDT | 0.378 | 0.536 | 0.184 | 0.068 | 0.334 | 1.363 | 0.454 | S M C |
| 5. | RCO | 0.376 | 0.507 | 0.155 ② | 0.066 | 0.400 | 1.704 | 0.384 | S M G |
| 6. | DRT | 0.356 | 0.519 | 0.201 | 0.062 | 0.321 | 1.503 | 0.426 | D M G |
| 7. | DeepSTRCF | 0.345 | 0.523 | 0.215 | 0.063 | 0.418 | 1.817 | 0.436 | D M G |
| 8. | CPT | 0.339 | 0.506 | 0.239 | 0.081 | 0.479 | 1.358 | 0.379 | D M G |
| 9. | SA_Siam_R | 0.337 | 0.566 ② | 0.258 | 0.337 ② | 0.566 ② | 0.258 ① | 0.429 | D P G |
| 10. | DLSTpp | 0.325 | 0.543 | 0.224 | 0.125 | 0.514 | 0.824 | 0.495 ① | S M G |
| 11. | LSART | 0.323 | 0.495 | 0.218 | 0.055 | 0.386 | 1.971 | 0.437 | S M G |
| 12. | SRCT | 0.310 | 0.520 | 0.290 | 0.059 | 0.331 | 1.765 | 0.400 | D M C |
| 13. | CFTR | 0.300 | 0.505 | 0.258 | 0.062 | 0.319 | 1.601 | 0.375 | D M G |
| 14. | CPT_fast | 0.296 | 0.520 | 0.290 | 0.152 | 0.515 | 0.726 | 0.392 | D M G |
| 15. | DeepCSRDCF | 0.293 | 0.489 | 0.276 | 0.062 | 0.399 | 1.644 | 0.393 | S M G |
| 16. | SiamVGG | 0.286 | 0.531 | 0.318 | 0.275 | 0.531 | 0.337 | 0.428 | D P G |
| 17. | SA_Siam_P | 0.286 | 0.533 | 0.337 | 0.286 ③ | 0.533 ③ | 0.342 | 0.406 | D P G |
| 18. | CFCF | 0.282 | 0.511 | 0.286 | 0.059 | 0.326 | 1.648 | 0.380 | D M G |
| 19. | ECO | 0.280 | 0.484 | 0.276 | 0.078 | 0.449 | 1.466 | 0.402 | D M G |
| 20. | MCCT | 0.274 | 0.532 | 0.318 | 0.061 | 0.359 | 1.742 | 0.422 | D M C |
| 21. | CCOT | 0.267 | 0.494 | 0.318 | 0.058 | 0.326 | 1.461 | 0.390 | D M G |
| 22. | csrtpp | 0.263 | 0.466 | 0.318 | 0.263 | 0.466 | 0.318 | 0.324 | D C G |
| 23. | LWDNTthi | 0.261 | 0.462 | 0.332 | 0.262 | 0.463 | 0.342 | 0.328 | D P G |
| 24. | LWDNTm | 0.261 | 0.455 | 0.323 | 0.261 | 0.455 | 0.323 | 0.352 | S P G |
| 25. | R_MCPF | 0.257 | 0.513 | 0.397 | 0.064 | 0.329 | 1.391 | 0.457 | S M G |
| 26. | FSAN | 0.256 | 0.554 ③ | 0.356 | 0.065 | 0.312 | 1.377 | 0.466 ③ | S M G |
| 27. | CSRDCF | 0.256 | 0.491 | 0.356 | 0.099 | 0.477 | 1.054 | 0.342 | D C C |
| 28. | DCFCF | 0.249 | 0.485 | 0.342 | 0.080 | 0.321 | 0.665 | 0.337 | D M C |
| 29. | UpdateNet | 0.244 | 0.518 | 0.454 | 0.209 | 0.517 | 0.534 | 0.358 | D M G |
| 30. | MBSiam | 0.241 | 0.529 | 0.443 | 0.238 | 0.529 | 0.440 | 0.413 | S P G |
| 31. | ALAL | 0.232 | 0.533 | 0.475 | 0.067 | 0.404 | 1.667 | 0.405 | S P G |
| 32. | CSTEM | 0.226 | 0.467 | 0.412 | 0.239 | 0.472 | 0.379 | 0.316 | S C C |
| 33. | BoVW_CFT | 0.224 | 0.500 | 0.450 | 0.063 | 0.331 | 1.615 | 0.373 | D M C |
| 34. | C3DT | 0.209 | 0.522 | 0.496 | 0.067 | 0.322 | 1.330 | 0.440 | D P G |
| 35. | RSECF | 0.206 | 0.470 | 0.501 | 0.074 | 0.414 | 1.569 | 0.319 | D M G |
| 36. | DSiam | 0.196 | 0.512 | 0.646 | 0.129 | 0.503 | 0.979 | 0.353 | D M G |
| 37. | KFebT | 0.195 | 0.474 | 0.674 | 0.195 | 0.475 | 0.670 | 0.221 | D C C |
| 38. | MEEM | 0.192 | 0.463 | 0.534 | 0.072 | 0.407 | 1.592 | 0.328 | S M C |
| 39. | SiamFC | 0.188 | 0.503 | 0.585 | 0.182 | 0.502 | 0.604 | 0.345 | D M G |
| 40. | STST | 0.187 | 0.464 | 0.621 | 0.156 | 0.466 | 0.763 | 0.297 | S P G |
| 41. | DCFNet | 0.182 | 0.470 | 0.543 | 0.180 | 0.471 | 0.548 | 0.327 | D M G |
| 42. | DensSiam | 0.174 | 0.462 | 0.688 | 0.174 | 0.462 | 0.688 | 0.305 | D P G |

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 43. | SAPKLTF | 0.171 | 0.488 | 0.613 | 0.117 | 0.481 | 0.946 | 0.352 | D C C |
| 44. | Staple | 0.169 | 0.530 | 0.688 | 0.170 | 0.530 | 0.688 | 0.335 | D M C |
| 45. | ASMS | 0.169 | 0.494 | 0.623 | 0.167 | 0.492 | 0.632 | 0.337 | D C C |
| 46. | ANT | 0.168 | 0.464 | 0.632 | 0.059 | 0.403 | 1.737 | 0.279 | D M C |
| 47. | HMMTxD | 0.168 | 0.506 | 0.815 | 0.073 | 0.416 | 1.564 | 0.330 | D C C |
| 48. | DPT | 0.158 | 0.486 | 0.721 | 0.126 | 0.483 | 0.899 | 0.315 | D C C |
| 49. | STBACF | 0.155 | 0.461 | 0.740 | 0.062 | 0.320 | 0.281 ③ | 0.245 | D M C |
| 50. | srdcf_deep | 0.154 | 0.492 | 0.707 | 0.057 | 0.326 | 1.756 | 0.321 | S M G |
| 51. | PBTS | 0.152 | 0.381 | 0.664 | 0.102 | 0.411 | 1.100 | 0.265 | S P C |
| 52. | DAT | 0.144 | 0.435 | 0.721 | 0.139 | 0.436 | 0.749 | 0.287 | D M C |
| 53. | LGT | 0.144 | 0.409 | 0.742 | 0.059 | 0.349 | 1.714 | 0.225 | S C C |
| 54. | RAnet | 0.141 | 0.449 | 0.744 | 0.133 | 0.477 | 0.805 | 0.303 | S P G |
| 55. | DFPReco | 0.138 | 0.473 | 0.838 | 0.049 | 0.312 | 0.286 | 0.269 | D M C |
| 56. | TRACA | 0.137 | 0.424 | 0.857 | 0.136 | 0.424 | 0.857 | 0.256 | D M G |
| 57. | KCF | 0.135 | 0.447 | 0.773 | 0.134 | 0.445 | 0.782 | 0.267 | D C C |
| 58. | FoT | 0.130 | 0.393 | 1.030 | 0.130 | 0.393 | 1.030 | 0.143 | D C C |
| 59. | srdcf_dif | 0.126 | 0.492 | 0.946 | 0.061 | 0.398 | 1.925 | 0.310 | D M G |
| 60. | SRDCF | 0.119 | 0.490 | 0.974 | 0.058 | 0.377 | 1.999 | 0.246 | S C C |
| 61. | MIL | 0.118 | 0.394 | 1.011 | 0.069 | 0.376 | 1.775 | 0.180 | S C C |
| 62. | BST | 0.116 | 0.272 | 0.881 | 0.053 | 0.271 | 1.620 | 0.149 | S C C |
| 63. | struck2011 | 0.097 | 0.418 | 1.297 | 0.093 | 0.419 | 1.367 | 0.197 | D C C |
| 64. | BDF | 0.093 | 0.367 | 1.180 | 0.093 | 0.367 | 1.180 | 0.145 | D C C |
| 65. | Matflow | 0.092 | 0.399 | 1.278 | 0.090 | 0.401 | 1.297 | 0.181 | S C C |
| 66. | MRSNCC | 0.082 | 0.330 | 1.506 | 0.060 | 0.328 | 2.088 | 0.112 | S M C |
| 67. | DSST | 0.079 | 0.395 | 1.452 | 0.077 | 0.396 | 1.480 | 0.172 | S C C |
| 68. | IVT | 0.076 | 0.400 | 1.639 | 0.065 | 0.386 | 1.854 | 0.130 | S C C |
| 69. | CPOINT | 0.070 | 0.308 | 1.719 | 0.057 | 0.290 | 1.901 | 0.115 | S M C |
| 70. | L1APG | 0.069 | 0.432 | 2.013 | 0.062 | 0.351 | 1.831 | 0.159 | S M C |
| 71. | FragTrack | 0.068 | 0.390 | 1.868 | 0.068 | 0.316 | 1.480 | 0.180 | S C C |
| 72. | Matrioska | 0.065 | 0.414 | 1.939 | 0.000 | 0.000 | 16.740 | 0.004 | S C C |

Table 2: The table shows expected average overlap (EAO), as well as accuracy and robustness raw values (A,R) for the baseline and the realtime experiments. For the unsupervised experiment the no-reset average overlap AO [91] is used. The last column contains implementation details (first letter: (D)eterministic or (S)tohastic, second letter: tracker implemented in (M)atlab, (C)++, or (P)ython, third letter: tracker is using (G)PU or only (C)PU).

The number of failures with respect to the visual attributes is shown in Figure 3. The overall top performers remain at the top of per-attribute ranks as well, but none of the trackers consistently outperforms all others with respect to each attribute. According to the median robustness and accuracy over each attribute (Table 1) the most challenging attributes in terms of failures are occlusion, illumination change and motion change, followed by camera motion and scale change. Occlusion is the most challenging attribute for tracking accuracy.

**The VOT-ST2018 winner identification** Top 10 trackers from the baseline experiment (Table 2) were selected to be re-run on the sequestered dataset. Despite significant effort, our team was unable to re-run DRT and SA_Siam_R due to library incompatibility errors in one case and significant system modifications requirements in the other. These two trackers were thus removed from the winner identification process on the account of the code provided not being results re-production-ready. The scores of the remaining trackers are shown in Table 3. The top tracker according to the EAO is MFT A.51 and is thus the VOT2018 short-term challenge winner.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



| | Tracker | EAO | A | R |
|---|---|---|---|---|
| 1. | MFT | 0.2518 ① | 0.5768 | 0.3105 ① |
| 2. | UPDT | 0.2469 ② | 0.6033 ② | 0.3427 ③ |
| 3. | RCO | 0.2457 ③ | 0.5707 | 0.3154 ② |
| 4. | LADCF | 0.2218 | 0.5499 | 0.3746 |
| 5. | DeepSTRCF | 0.2205 | 0.5998 ③ | 0.4435 |
| 6. | CPT | 0.2087 | 0.5773 | 0.4238 |
| 7. | SiamRPN | 0.2054 | 0.6277 ① | 0.5175 |
| 8. | DLSTpp | 0.1961 | 0.5833 | 0.4544 |

Table 3: The top eight trackers from Table 2 re-ranked on the VOT2018 sequestered dataset.

### 4.3 The VOT2018 short-term real-time sub-challenge results

The EAO scores and AR-raw plots for the real-time experiment are shown in Figure 4 and Figure 5. The top ten real-time trackers are SiamRPN A.35, SA_Siam_R A.60, SA_Siam_P A.59, SiamVGG A.63, CSRTPP A.25, LWDNTm A.41, LWDNTthi A.42, CSTEM A.9, MBSiam A.48 and UpdateNet A.1. Eight of these (SiamRPN, SA_Siam_R, SA_Siam_P, SiamVGG, LWDNTm, LWDNTthi, MBSiam, UpdateNet) are extensions of the Siamese architecture SiamFC [6]. These trackers apply pre-traind CNN features that maximize correlation localization accuracy and require a GPU. But since feature extraction as well as correlation are carried out on the GPU, they achieve significant speed in addition to extraction of highly discriminative features. The remaining two trackers (CSRTPP and CSTEM) are extensions of the CSRDCF [53] – a correlation filter with boundary constraints and segmentation for identifying reliable target pixels. These two trackers apply hand-crafted features, i.e., HOG and Colornames.

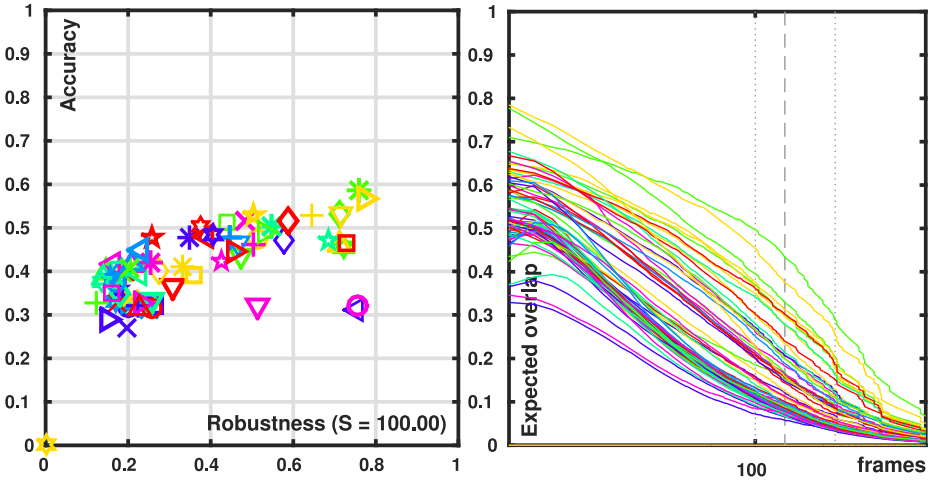


Fig. 4: The AR plot (left) and the EAO curves (right) for the VOT2017 realtime experiment.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



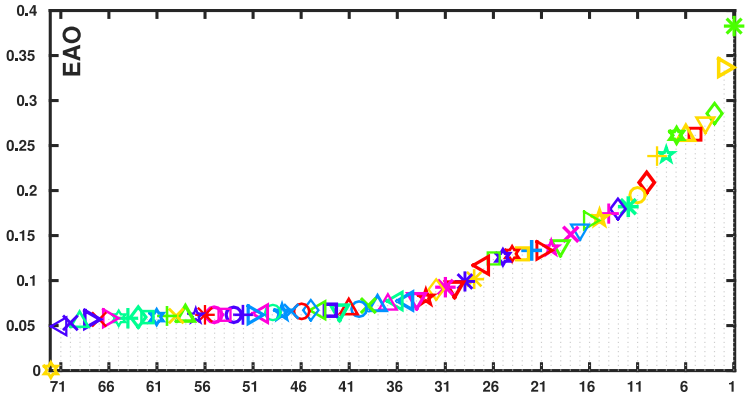


Fig. 5: The EAO plot (right) for the realtime experiment.

**The VOT-RT2018 winner identification** The winning real-time tracker of the VOT2018 is the Siamese region proposal network SiamRPN [48] (A.35). The tracker is based on a Siamese subnetwork for feature extraction and a region proposal subnetwork which includes a classification branch and a regression branch. The inference is formulated as a local one-shot detection task.

## 5 The VOT2018 long-term challenge results

The VOT2018 LT challenge received 11 valid entries. The VOT2018 committee contributed additional 4 baselines, thus 15 trackers were considered in the VOT2018 LT challenge. In the following we briefly overview the entries and provide the references to original papers in the Appendix B where available.

Some of the submitted trackers were in principle $ST_0$ trackers. But the submission rules required exposing a target localization/presence certainty score which can be used by thresholding to form a target presence classifier. In this way, these trackers were elevated to $LT_0$ level according to the ST-LT taxonomy from Section 3.1. Five trackers were from the $ST_0$ (elevated to $LT_0$) class: SiamVGG B.15, SiamFC B.5, ASMS B.11, FoT B.3 and SLT B.14. Ten trackers were from $LT_1$ class: DaSiam_LT B.2, MMLT B.1, PTAVplus B.10, MBMD B.8, SAPKLTF B.12, LTSINT B.7, SYT B.13, SiamFCDet B.4, FuCoLoT B.6, HMMTxD B.9.

Ten trackers applied CNN features (nine of these in Siamese architecture) and four trackers applied DCFs. Six trackers never updated *the short-term component* (DaSiam_LT, SYT, SiamFCDet, SiamVGG, SiamFC and SLT), four updated the component only when confident (MMLT, SAPKLTF, LTSINT, FuCoLoT), two applied exponential forgetting (HMMTxD, ASMS), two applied updates at fixed intervals (PTAVplus, MBMD) and one applied robust partial updates (FoT). Seven trackers never updated *the long-term component* (DaSiam_LT, MBMD, SiamFCDet, HMMTxD, SiamVGG, SiamFC, SLT), and six updated the model only when confident (MMLT, PTAVplus, SAPKLTF, LTSINT, SYT, FuCoLoT).

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



| | Tracker | F-score | Pr | Re | ST/LT | Frames (Success) |
|---|---|---|---|---|---|---|
| 1. | MBMD | 0.610 ① | 0.634 ② | 0.588 ① | $LT_1$ | 1 (100%) |
| 2. | DaSiam_LT | 0.607 ② | 0.627 ③ | 0.588 ② | $LT_1$ | - (0%) |
| 3. | MMLT | 0.546 ③ | 0.574 | 0.521 ③ | $LT_1$ | 0 (100%) |
| 4. | LTSINT | 0.536 | 0.566 | 0.510 | $LT_1$ | 2 (100%) |
| 5. | SYT | 0.509 | 0.520 | 0.499 | $LT_1$ | 0 (43%) |
| 6. | PTAVplus | 0.481 | 0.595 | 0.404 | $LT_1$ | 0 (11%) |
| 7. | FuCoLoT | 0.480 | 0.539 | 0.432 | $LT_1$ | 78 (97%) |
| 8. | SiamVGG | 0.459 | 0.552 | 0.393 | $ST_0 \rightarrow LT_0$ | - (0%) |
| 9. | SLT | 0.456 | 0.502 | 0.417 | $ST_1 \rightarrow LT_0$ | 0 (100%) |
| 10. | SiamFC | 0.433 | 0.636 ① | 0.328 | $ST_0 \rightarrow LT_0$ | - (0%) |
| 11. | SiamFCDet | 0.401 | 0.488 | 0.341 | $LT_1$ | 0 (83%) |
| 12. | HMMTxD | 0.335 | 0.330 | 0.339 | $LT_1$ | 3 (91%) |
| 13. | SAPKLTF | 0.323 | 0.348 | 0.300 | $LT_0$ | - (0%) |
| 14. | ASMS | 0.306 | 0.373 | 0.259 | $ST_0 \rightarrow LT_0$ | - (0%) |
| 15. | FoT | 0.119 | 0.298 | 0.074 | $ST_0 \rightarrow LT_0$ | 0 (6%) |

Table 4: List of trackers that participated in the VOT2018 long-term challenge along with their performance scores (F-score, Pr, Re), ST/LT categorization and results of the re-detection experiment in the last column with the average number of frames required for re-detection (Frames) and the percentage of sequences with successful re-detection (Success).

Results of the re-detection experiment are summarized in the last column of Table 4. MMLT, SLT, MBMD, FuCoLoT and LTSINT consistently re-detect the target while SiamFCDet succeeded in all but one sequence. Some trackers (SYT, PTAVplus) were capable of re-detection in only a few cases, which indicates a potential issue with the detector. All these eight trackers pass the re-detection test and are classified as $LT_1$ trackers. Trackers DaSiam_LT, SAPKLTF, SiamVGG and SiamFC did not pass the test, which means that they do not perform image-wide re-detection, but only re-detect in a extended local region. These trackers are classified as $LT_0$.

The overall performance is summarized in Figure 6. The highest ranked tracker is the MobileNet-based tracking by detection algorithm (MBMD), which applies a bounding box regression network and an MDNet-based verifier [64]. The bounding box regression network is trained on ILSVRC 2015 video detection dataset and ILSVRC 2014 detection dataset is used to train a regression to any object in a search region by ignoring the classification labels. The bounding box regression result is verified by MDNet [64]. If the score of regression module is below a threshold, the MDNet localizes the target by a particle filter. The MDNet is updated online, while the bounding box regression network is not updated.

The second highest ranked tracker is DaSiam_LT – an $LT_1$ class tracker. This tracker is an extension of a Siamese Region Proposal Network (SiamRPN) [48]. The original SiamRPN cannot recover a target after it re-appears, thus the extension implements an effective global-to-local search strategy. The search region size is gradually grown at a constant rate after target loss, akin to [55].

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



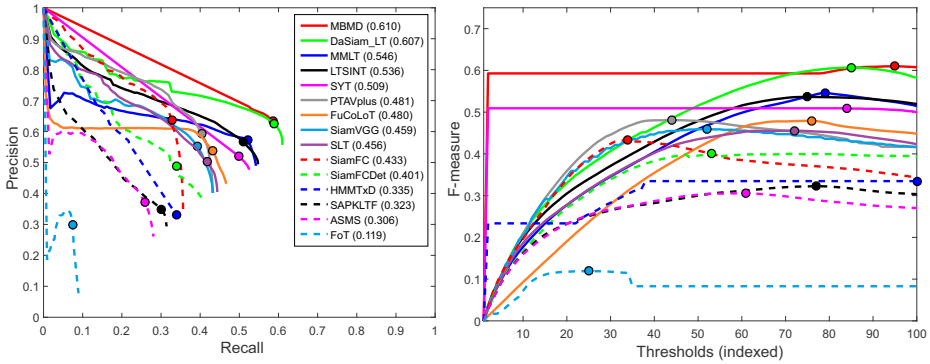


Fig. 6: Long-term tracking performance. The average tracking precision-recall curves (left), the corresponding F-score curves (right). Tracker labels are sorted according to maximum of the F-score.

Distractor-aware training and inference are also added to implement a high-quality tracking reliability score.

Figure 7 shows tracking performance with respect to nine visual attributes from Section 3.2. The most challenging attributes are fast motion, out of view, aspect ratio change and full occlusion.

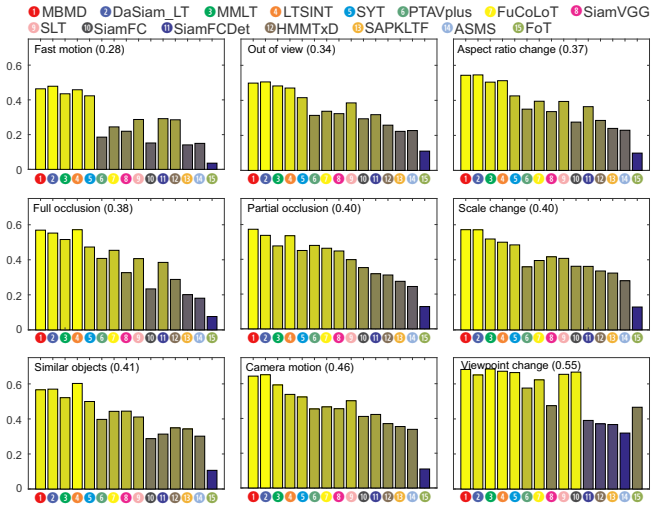


Fig. 7: Maximum F-score averaged over overlap thresholds for the visual attributes. The most challenging attributes are fast motion, out of view, aspect ratio change and full occlusion.

**The VOT-LT2018 winner identification** According to the F-score, MBMD (F-score=0,610) is slightly ahead of DaSiam_LT (F-score=0,607). The trackers

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



reach approximately the same tracking recall (0,588216 for MBMD vs 0,587921 for DaSiam_LT), which implies a comparable target re-detection success. But MBMD has a greater tracking precision which implies better target localization capabilities. Overall, the best tracking precision is obtained by SiamFC, while the best tracking recall is obtained by MBMD. According to the VOT winner rules, the VOT2018 long-term challenge winner is therefore MBMD B.8.

## 6 Conclusion

Results of the VOT2018 challenge were presented. The challenge is composed of the following three sub-challenges: the main VOT2018 short-term tracking challenge (VOT-ST2018), the VOT2018 real-time short-term tracking challenge (VOT-RT2018) and VOT2018 long-term tracking challenge (VOT-LT2018), which is a new challenge introduced this year.

The overall results of the challenges indicate that discriminative correlation filters and deep networks remain the dominant methodologies in visual object tracking. Deep features in DCFs and use of CNNs as classifiers in the trackers have been recognized as efficient tracking ingredients already in VOT2015. But their use among top performers has become wide-spread over the following years. In contrast to previous years we observe a wider use of localization-trained CNN features and CNN trackers based on Siamese architectures. Bounding box regression is being used in trackers more frequently than in previous challenges as well.

The top performer on the VOT-ST2018 *public dataset* is LADCF (A.39) – a regularized discriminative correlation filter trained on a low-dimensional projection of ResNet50, HOG and Colornames features. The top performer on the sequestered dataset and the VOT-ST2018 challenge winner is MFT (A.51) – a continuous convolution discriminative correlation filter with per-channel independently trained localization learned features. This tracker uses ResNet-50, SE-ResNet-50, HOG and Colornames.

The top performer and the winner of the VOT-RT2018 challenge is SiamRPN (A.35) – a Siamese region proposal network. The tracker requires a GPU, but otherwise has the best tradeoff between robustness and processing speed. Note that nearly all top ten trackers on realtime challenge applied Siamese nets (two applied DCFs and run on CPU). The dominant methodology in real-time tracking therefore appears to be Siamese CNNs.

The top performer and the winner of the VOT-LT2018 challenge is MBMD (B.8) – a bounding box regression network with MDNet [64] for regression verification and localization upon target loss. This tracker is from $LT_1$ class, identifies a potential target loss, performs target re-detection and applies conservative updates of the visual model.

The VOT primary objective is to establish a platform for discussion of tracking performance evaluation and contributing to the tracking community with verified annotated datasets, performance measures and evaluation toolkits. The

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



VOT2018 was a sixth effort toward this, following the very successful VOT2013, VOT2014, VOT2015, VOT2016 and VOT2017.

## Acknowledgements

This work was supported in part by the following research programs and projects: Slovenian research agency research programs P2-0214, P2-0094, Slovenian research agency project J2-8175. Jiři Matas and Tomáš Vojíř were supported by the Czech Science Foundation Project GACR P103/12/G084. Michael Felsberg and Gustav Häger were supported by WASP, VR (EMC2), SSF (SymbiCloud), and SNIC. Roman Pflugfelder and Gustavo Fernández were supported by the AIT Strategic Research Programme 2017 Visual Surveillance and Insight. The challenge was sponsored by Faculty of Computer Science, University of Ljubljana, Slovenia.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



# A VOT2018 short-term challenge tracker descriptions

In this appendix we provide a short summary of all trackers that were considered in the VOT2018 short-term challenges.

## A.1 Adaptive object update for tracking (UpdateNet)

*L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, F. S. Khan*
{*lichao, agonzalez, joost*}*@cvc.uab.es, fahad.khan@liu.se*

UpdateNet tracker uses an update network to update the tracked object appearance during tracking. Since the object appearance constantly changes as the video progresses, some update mechanism is necessary to maintain an accurate model of the object appearance. The traditional correlation tracker updates the object appearance by using a fixed update rule based on a single hyperparameter. This approach, however, cannot effectively adapt to the specific update requirement necessary for every particular situation. UpdateNet extends the correlation tracker of SiamFC [6] to include a network component specially trained to update the object appearance which is an advantage with respect to the traditional fixed rule update used for tracking.

## A.2 Anti-decay LSTM with Adversarial Learning Tracker (ALAL)

*F. Zhao, Y. Wu, J. Wang, M. Tang*
{*fei.zhao, jqwang, tangm*}*@nlpr.ia.ac.cn, ywu.china@gmail.com*

The ALAL tracker contains two CNNs: a regression CNN and a classification CNN. For each search patch, the former CNN predicts a response map which reflects the location of the target. The latter CNN distinguishes the target from the candidates. A modified LSTM which is trained by the adversarial learning is also added on the former network. The modified LSTM can extract the features of the target in long-term without the decay of the feature.

## A.3 ANT (ANT)

*Submitted by VOT Committee*

The ANT tracker is a conceptual increment to the idea of multi-layer appearance representation that is first described in [82]. The tracker addresses the problem of self-supervised estimation of a large number of parameters by introducing controlled graduation in estimation of the free parameters. The appearance of the object is decomposed into several sub-models, each describing the target at a different level of detail. The sub models interact during target localization and, depending on the visual uncertainty, serve for cross-sub-model supervised updating. The reader is referred to [84] for details.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### A.4 Bag-of-Visual-Words based Correlation Filter Tracker (BoVW_CFT)

*P. M. Raju, D. Mishra, G. R. K. S. Subrahmanyam*
{*priyamariyam123, vr.dkmishra*}*@gmail.com, rkg@iittp.ac.in*

The BoVW-CFT is a classifier-based generic technique to handle tracking uncertainties in correlation filter trackers. The method is developed using ECO [15] as the base correlation tracker. The classifier operates on Bag of Visual Words (BoVW) features and SVM with training, testing and update stages. For each tracking uncertainty, two output patches are obtained, one each from the base tracker and the classifier. The final output patch is the one with highest normalized cross-correlation with the initial target patch.

### A.5 Best Displacement Flow (BDF)

*M. E. Maresca, A. Petrosino*
*mariomaresca@hotmail.it, alfredo.petrosino@uniparthenope.it*

Tracker BDF is based on the idea of Flock of Trackers [86] in which a set of local tracker responses are robustly combined to track the object. The reader is referred to [58] for details.

### A.6 Best Structured Tracker (BST)

*F. Battistone, A. Petrosino, V. Santopietro*
{*francesco.battistone, alfredo.petrosino, vincenzo.santopietro*}*@uniparthenope.it*

BST is based on the idea of Flock of Trackers [86]: a set of five local trackers tracks a little patch of the original target and then the tracker combines their information in order to estimate the resulting bounding box. Each local tracker separately analyzes the Haar features extracted from a set of samples and then classifies them using a structured Support Vector Machine as Struck [28]. Once having predicted local target candidates, an outlier detection process is computed by analyzing the displacements of local trackers. Trackers that have been labeled as outliers are reinitialized. At the end of this process, the new bounding box is calculated using the Convex Hull technique. For more detailed information, please see [5].

### A.7 Channel pruning for visual tracking (CPT)

*M. Che, R. Wang, Y. Lu, Y. Li, H. Zhi, C. Xiong*
*cmq_mail@163.com,* {*1573112241, 1825650885*}*@qq.com,*
*liyan1994626@126.com, 1462714176@qq.com, xczkiong@163.com*

In order to improve the tracking speed, the tracker CPT is proposed. The tracker introduces an effective channel pruning based VGG network to fast extract the deep convolutional features. In this way, it can obtain deeper convolutional features for better representations of various objects' variations without

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



worrying about the speed of suppression. To further reduce the redundancy features, the Average Feature Energy Ratio is proposed to extract effective convolutional channel of the selected deep convolution layer and increase the tracking speed. The method also ameliorates the optimization process in minimizing the location error as adaptive iterative optimization strategy.

### A.8 Channel pruning for visual tracking (CPT_fast)

*M. Che, R. Wang, Y. Lu, Y. Li, H. Zhi, C. Xiong*
*cmq_mail@163.com, {1573112241, 1825650885}@qq.com,*
*liyan1994626@126.com, 1462714176@qq.com, xczkiong@163.com*

The fast CPT (called CPT_fast) method is based on CPT tracker A.7 and the DSST [12] method which is applied to estimate the tracking object's scale.

### A.9 Channels-weighted and Spatial-related Tracker with Effective response-map Measurement (CSTEM)

*Z. Zhang, Y. Li, J. Ren, J. Zhu*
*{zzheng1993, liyang89, zijinxuxu, jkzhu}@zju.edu.cn*

Motivated by CSRDCF tracker [53], CSTEM has designed an effective measurement function to evaluate the quality of filter response. As a theoretical guarantee of effectiveness, CSTEM tracker scheme chooses different filter models according to the different scenarios using the measurement function. Moreover, a sophisticated strategy is employed to detect occlusion, and then decide how to update the filter models in order to alleviate the drifting problem. In addition, CSTEM takes advantage of both log-polar approach [50] and pyramid-like method [12] to accurately estimate the scale changes of the tracking target. For the detailed information, please see [99].

### A.10 Combined Point Tracker (CPOINT)

*A. G. Perera, Y. W. Law, J. Chahl*
*asanka.perera@mymail.unisa.edu.au, {yeewei.law, javaan.chahl}@unisa.edu.au*

CPOINT tracker combines 3 different trackers to predict and correct the target location and size. In the first level, four types of key-point features (SURF, BRISK, KAZE and FAST) are used to localize and scale up or down the bounding box of the target. The size and the location of the initial estimation is averaged out with another level of corner point tracker which also uses optical flow. Predictions with insufficient image details are handled by a third level histogram-based tracker.

### A.11 Continuous Convolution Operator Tracker (CCOT)

*Submitted by VOT Committee*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



C-COT learns a discriminative continuous convolution operator as its tracking model. C-COT poses the learning problem in the continuous spatial domain. This enables a natural and efficient fusion of multi-resolution feature maps, e.g. when using several convolutional layers from a pre-trained CNN. The continuous formulation also enables highly accurate localization by sub-pixel refinement. The reader is referred to [17] for details.

### A.12 Continuous Convolution Operators with Resnet features (RCO)

*Z. He, S. Bai, J. Zhuang*
{*he010103, baishuai*}*@bupt.edu.cn, junfei.zhuang@faceall.cn*

The RCO tracker is based on an extension of CFWCR [31]. A continuous convolution operator is used to fuse multi-resolution features synthetically, which improves the performance of correlation filter based tracker. Shallower and deeper features from convolution neural network focus on different target information. In order to improve the cooperative solving method and make full use of diverse features a multi-solution is proposed. To predict the target location RCO optimally fuses the obtained multi-solutions. RCO tracker uses CNN features extracted from Resnet50.

### A.13 Convolutional Features for Correlation Filters (CFCF)

*E. Gundogdu, A. Alatan*
*erhan.gundogdu@epfl.ch, alatan@metu.edu.tr*

The tracker CFCF is based on the feature learning study in [26] and the correlation filter based tracker C-COT [17]. The proposed tracker employs a fully convolutional neural network (CNN) model trained on ILSVRC15 video dataset [71] by the learning framework introduced in [26] which is designed for correlation filter [12]. To learn features, convolutional layers of VGG-M-2048 network [11] trained on [19] are applied. An extra convolutional layer is used for fine-tuning on ILSVRC15 dataset. The first, fifth and sixth convolutional layers of the learned network, HOG [63] and Colour Names (CN) [89] are integrated to the C-COT tracker [17].

### A.14 Correlation Filter with Regressive Scale Estimation (RSECF)

*L. Chu, H. Li*
{*lt.chu, hy.li*}*@siat.ac.cn*

RSECF addresses the problems of poor scale estimation in state of art DCF trackers by learning separate discriminative correlation filters for translation estimation and bounding box regression for scale estimation. The scale filter is learned online using the target appearance sampled at a set of different aspect ratios. Contrary to standard approaches, RSECF directly searches for continuous scale space, which can predict any scale without being limited by manually specified number of scales. RSECF generalizes the original single-channel bounding

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



box regression to multi-channel situations, which allows for more efficient employment of multi-channel features. The correlation filter is ECOhc [15] without fDSST [16],which locates the target position.

### A.15 Correlation Filter with Temporal Regression (CFTR)

*L. Rout, D. Mishra, R. K. Gorthi*
*liturout1997@gmail.com, deepak.mishra@iist.ac.in, rkg@iittp.ac.in*

CFTR tracker proposes a different approach to regress in the temporal domain based on the Tikhonov regularization. CFTR tracker applies a weighted aggregation of distinctive visual features and feature prioritization with entropy estimation in a recursive fashion. A statistics based ensembler approach is proposed for integrating the conventionally driven spatial regression results (such as from CFCF [26]), and the proposed temporal regression results to accomplish better tracking.

### A.16 Correlation Tracking via Joint Discrimination and Reliability Learning (DRT)

*C. Sun, Y. Zhang, Y. Sun, D. Wang, H. Lu*
{*waynecool, zhangyunhua, rumsyx*}*@mail.dlut.edu.cn,*
{*wdice, lhchuan*}*@dlut.edu.cn*

DRT uses a novel CF-based optimization problem to jointly model the discrimination and reliability information. First, the tracker treats the filter as the element-wise product of a base filter and a reliability term. The base filter is aimed to learn the discrimination information between the target and backgrounds, and the reliability term encourages the final filter to focus on more reliable regions. Second, the DRT tracker introduces a local response consistency regular term to emphasize equal contributions of different regions and avoid the tracker being dominated by unreliable regions. The tracker is based on [77].

### A.17 CSRDCF with the integration of CNN features and handcrafted features (DeepCSRDCF)

*Z. He*
*he010103@bupt.edu.cn*

DeepCSRDCF adopts CSRDCF tracker [53] as the baseline approach. CNN features are integrated into hand-crafted features, which boosts the performance compared to the baseline tracker CSRDCF. To avoid the model drift, an adaptive learning rate is applied.

### A.18 DCFNET: Discriminant Correlation Filters Network for Visual Tracking (DCFNet)

*J. Li, Q. Wang, W. Hu*
*jli24@outlook.com, wangqiang2015@ia.ac.cn, wmhu@nlpr.ia.ac.cn*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



DCFNet is a tracker with the end-to-end lightweight network architecture, which learned the convolutional features and performed the correlation tracking process simultaneously. Specifically, DCF is treated as a special correlation filter layer added in a Siamese network. The back-propagation through the network is derived by defining the network output as the probability heat-map of the object location. Since the derivation is still carried out in Fourier frequency domain, the efficiency property of DCF is preserved. For more detailed information on this tracker, please see reference [88].

### A.19 Deep Enhanced Spatially Regularized Discriminative Correlation Filter (srdcf_deep)

*J. Rodríguez Herranz, V. Štruc, K. Grm*
*j.rodriguezherranz@gmail.com, {vitomir.struc, klemen.grm}@fe.uni-lj.si*

The Deep Enhanced Spatially Regularized Discriminative Correlation Filter (srdcf_deep) is based on the E-SRDCF tracker incorporating the constrained correlation filter from [13] and a motion model based on frame differences. While E-SRDCF uses only hand-crafted features (HOGs, colour names and grey-scale images), DE-SRDCF also exploits learned CNN-based features. Specifically, the CNN model used for feature extraction is an auto-encoder with a similar architecture as VGG-m [11]. The features used are taken from the first and fifth convolutional layer. More information on DE-SRDCF tracker can be found in [33].

### A.20 DeepSTRCF (DeepSTRCF)

*W. Zuo, F. Li, X. Wu, C. Tian, M.-H. Yang*
*cswmzuo@gmail.com, fengli_hit@hotmail.com, xhwu.cpsl.hit@gmail.com, tcoperator@163.com, mhyang@ucmerced.edu*

DeepSTRCF implements a variant of STRCF tracker [49] with deep CNN features. STRCF addresses the computational inefficiency problem of SRDCF tracker from two aspects: (i) a temporal regularization term to remove the need of formulation on large training sets, and (ii) an ADMM algorithm to solve the STRCF model efficiently. Therefore, it can provide more robust models and much faster solutions than SRDCF thanks to the online Passive-Aggressive learning and ADMM solver, respectively.

### A.21 Deformable part correlation filter tracker (DPT)

*Submitted by VOT Committee*

DPT is a part-based correlation filter composed of a coarse and mid-level target representations. Coarse representation is responsible for approximate target localization and uses HOG as well as colour features. The mid-level representation is a deformable parts correlation filter with fully-connected parts topology and applies a novel formulation that threats geometric and visual properties

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



within a single convex optimization function. The mid level as well as coarse level representations are based on the kernelized correlation filter from [32]. The reader is referred to [54] for details.

### A.22 Dense Contrastive Features for Correlation Filters (DCFCF)

*J. Spencer Martin, R. Bowden, S. Hadfield*
{*jaime.spencer, r.bowden, s.hadfield*}*@surrey.ac.uk*

Dense Contrastive Features for Correlation Filters (DCFCF) extends on previous work based on correlation filters applied to feature representations of the tracked object. A new type of dense feature descriptors is introduced which is specifically trained for the comparison of unknown objects. These generic comparison features lead to a more robust representation of a priori unknown objects, largely increasing the resolution compared to intermediate layers, whilst maintaining a reasonable dimensionality. This results in a slight increase in performance, along with a higher resistance to occlusions or missing targets.

### A.23 Densely connected Siamese architecture for robust visual tracking (DensSiam)

*M. Abdelpakey, M. Shehata*
{*mha241, mshehata*}*@mun.ca*

DensSiam is a new Siamese architecture for object tracking. It uses the concept of dense layers and connects each dense layer to all layers in a feed-forward fashion with a similarity-learning function. DensSiam uses non-local features to represent the appearance model in such a way that allows the deep feature map to be robust to appearance changes. DensSiam allows different feature levels (e.g. low level and high-level features) to flow through the network layers without vanishing gradients and improves the generalization capability [1].

### A.24 Discriminative Correlation Filter with Channel and Spatial Reliability (CSRDCF)

*Submitted by VOT Committee*

The CSRDCF [53] improves discriminative correlation filter trackers by introducing two concepts: spatial reliability and channel reliability. It uses colour segmentation as spatial reliability to adjust the filter support to the part of the object suitable for tracking. The channel reliability reflects the discriminative power of each filter channel. The tracker uses HoG and colour-names features.

### A.25 Discriminative Correlation Filter with Channel and Spatial Reliability - C++ (csrtpp)

*Submitted by VOT Committee*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



The csrtpp tracker is the C++ implementation of the Discriminative Correlation Filter with Channel and Spatial Reliability (CSR-DCF) tracker A.24.

### A.26 Discriminative Scale Space Tracker (DSST)

*Submitted by VOT Committee*

The Discriminative Scale Space Tracker (DSST) [12] extends the Minimum Output Sum of Squared Errors (MOSSE) tracker [9] with robust scale estimation. The DSST additionally learns a one-dimensional discriminative scale filter, that is used to estimate the target size. For the translation filter, the intensity features employed in the MOSSE tracker is combined with a pixel-dense representation of HOG-features.

### A.27 Distractor-aware Tracking (DAT)

*H. Possegger*
*possegger@icg.tugraz.at*

The Tracker DAT [68] is an appearance-based tracking-by-detection approach. It relies on a generative model using colour histograms to distinguish the object from its surroundings. Additionally, a distractor-aware model term suppresses visually similar (i.e. distracting) regions whenever they appear within the field-of-view, thus reducing tracker drift.

### A.28 DLSTpp: Deep Location-Specific Tracking++ (DLSTpp)

*L. Yang*
*lingxiao.yang717@gmail.com*

The DLSTpp is a tracker based on DLST tracker which decomposes the tracking problem into a localization and a classification task. The localization is achieved by ECOhc. The classification network is the same as MDNet, but their weights are fine-tuned on ImageNet VID dataset.

### A.29 Dynamic Fusion of Part Regressors for Correlation Filter-based Visual Tracking (DFPReco)

*A. Memarmoghadam, P. Moallem*
{*a.memarmoghadam, p_moallem*}*@eng.ui.ac.ir*

Employing both global and local part-wise appearance models, a robust tracking algorithm based on weighted fusion of several CF-based part regressors is proposed. Importance weights are dinamically assigned to each part via solving a multi-linear ridge regression optimization problem towards achieving a more discriminative target-level confidence map. Additionally it is presented an accurate size estimation method that jointly provides object scale and aspect ratio by analyzing relative deformation cost of importance pair-wise parts. A

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



single-patch ECO tracker [15] (but without object scale mechanism) is applied as baseline approach for each part which expeditiously makes track of target object parts.

### A.30 Dynamic Siamese Network based Tracking (DSiam)

*Q. Guo, W. Feng*
{*tsingqguo, wfeng*}*@tju.edu.cn*

DSiam [27] locates an interested target by matching an online updated template with a suppressed search region. This is achieved by adding two transformations to the two branches of a pretained network that can be SiamFC, VGG19, VGG16, etc. The two transformations can be efficiently online learned in frequency domain. Instead of using the pretrained network in [27], the presented tracker uses the network introduced in [81] to extract deep features.

### A.31 ECO (ECO)

*Submitted by VOT Committee*

ECO addresses the problems of computational complexity and over-fitting in state of the art DCF trackers by introducing: (i) a factorized convolution operator, which drastically reduces the number of parameters in the model; (ii) a compact generative model of the training sample distribution, that significantly reduces memory and time complexity, while providing better diversity of samples; (iii) a conservative model update strategy with improved robustness and reduced complexity. The reader is referred to [15] for more details.

### A.32 Enhanced Spatially Regularized Discriminative Correlation Filter (srdcf_dif)

*J. Rodríguez Herranz, V. Štruc, K. Grm*
*j.rodriguezherranz@gmail.com,* {*vitomir.struc, klemen.grm*}*@fe.uni-lj.si*

The Enhanced Spatially Regularized Discriminative Correlation Filter (srdcf_dif) is based on the constrained correlation filter formulation from [13], but incorporates an additional motion model to improve tracking performance. The motion model takes again the form of a constrained correlation filter, but is computed over frame differences instead of static frames. The standard SRDCF tracker and motion model are combined using a weighted sum over the correlation outputs. Both E-SRDCF parts exploit HOG, colour names and grey scale image features during filter construction. For more details the reader is referred to [33].

### A.33 Flock of Trackers (FoT)

*Submitted by VOT Committee*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



The Flock of Trackers (FoT) is a tracking framework where the object motion is estimated from the displacements or, more generally, transformation estimates of a number of local trackers covering the object. Each local tracker is attached to a certain area specified in the object coordinate frame. The local trackers are not robust and assume that the tracked area is visible in all images and that it undergoes a simple motion, e.g. translation. The FoT object motion estimate is robust if it is from local tracker motions by a combination which is insensitive to failures.

### A.34 Fully-Convolutional Siamese Network (SiamFC)

*L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi, P. Torr*
*{luca.bertinetto, joao.henriques, andrea.vedaldi, philip.torr}@eng.ox.ac.uk, jack.valmadre@gmail.com*

SiamFC applies a fully-convolutional deep Siamese conv-net to locate the best match for an exemplar image within a larger search image. The deep conv-net is trained offline on video detection datasets to address a general similarity learning problem.

### A.35 High Performance Visual Tracking with Siamese Region Proposal Network (SiamRPN)

*Q. Wang, Z. Zhu, B. Li, W. Wu, W. Hu, W. Zou*
*{wangqiang2015, zhuzheng2014}@ia.ac.cn, lbvictor2013@gmail.com, wuwei@sensetime.com, wmhu@nlpr.ia.ac.cn, wei.zou@ia.ac.cn*

The tracker SiamRPN consists of a Siamese sub-network for feature extraction and a region proposal sub-network including the classification branch and regression branch. In the inference phase, the proposed framework is formulated as a local one-shot detection task. The template branch of the Siamese sub-network is pre-computed while correlation layers are formulated as convolution layers to perform online tracking [48]. What is more, SiamRPN introduces an effective sampling strategy to control the imbalanced sample distribution and make the model focus on the semantic distractors [102].

### A.36 Incremental Learning for Robust Visual Tracking (IVT)

*Submitted by VOT Committee*

The idea of the IVT tracker [70] is to incrementally learn a low-dimensional sub-space representation, adapting on-line to changes in the appearance of the target. The model update, based on incremental algorithms for principal component analysis, includes two features: a method for correctly updating the sample mean, and a forgetting factor to ensure less modelling power is expended fitting older observations.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### A.37 Kalman Filter ensemble-based Tracker (KFebT)

*P. Senna, I. Drummond, G. Bastos*
*pedro.senna@ufms.br, isadrummond@unifei.edu.br, sousa@unifei.edu.br*

The tracker KFebT [72] fuses the result of two out-of-the box trackers, a mean-shift tracker that uses colour histogram (ASMS) [87] and a kernelized correlation filter (KCF) [32] by using a Kalman filter. Compared from last year submission, current version includes a partial feedback and an adaptive model update. Code available at https://github.com/psenna/KF-EBT.

### A.38 Kernelized Correlation Filter (KCF)

*Submitted by VOT Committee*

This tracker is a C++ implementation of Kernelized Correlation Filter [32] operating on simple HOG features and Colour Names. The KCF tracker is equivalent to a Kernel Ridge Regression trained with thousands of sample patches around the object at different translations. It implements multi-thread multi-scale support, sub-cell peak estimation and replacing the model update by linear interpolation with a more robust update scheme. Code available at https://github.com/vojirt/kcf.

### A.39 Learning Adaptive Discriminative Correlation Filter on Low-dimensional Manifold (LADCF)

*T. Xu, Z.-H. Feng, J. Kittler, X.-J. Wu*
*tianyang_xu@163.com, {z.feng, j.kittler}@surrey.ac.uk,*
*wu_xiaojun@jiangnan.edu.cn*

LADCF utilises adaptive spatial regularizer to train low-dimensional discriminative correlation filters [93]. A low-dimensional discriminative manifold space is designed by exploiting temporal consistency, which realises reliable and flexible temporal information compression, alleviating filter degeneration and preserving appearance diversity. Adaptive spatial regularization and temporal consistency are combined in an objective function, which is optimised by the augmented Lagrangian method. Robustness is further considered by integrating HOG, Colour Names and ResNet-50 features. For ResNet-50 features, data augmentation [8] is adopted using flip, rotation and blur. The tracker is implemented on MatLab running on the CPU.

### A.40 Learning Spatial-Aware Regressions for Visual Tracking (LSART)

*C. Sun, Y. Sun, S. Wang, D. Wang, H. Lu, M.-H. Yang*
*{waynecool, rumsyx, wwen9502}@mail.dlut.edu.cn,*
*{wdice, lhchuan}@dlut.edu.cn, mhyang@ucmerced.edu*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



The LSART tracker exploits the complementary kernelized ridge regression (KRR) and convolution neural network (CNN) for tracking. A weighted cross-patch similarity kernel for the KRR model is defined and the spatially regularized filter kernels for the CNN model is used. While the former focuses on the holistic target, the latter focuses on the small local regions. The distance transform is exploited to pool layers for the CNN model, which determines the reliability of each output channel. Three kinds of features are used in the proposed method: Conv4-3 of VGG-16, Hog, and Colour naming. The LSART tracker is based on [78].

### A.41 Lightweight Deep Neural Network for Visual Tracking (LWDNTm)

*H. Zhao, D. Wang, H. Lu*
*zhaohj@stumail.neu.edu.cn, {wdice, lhchuan}@dlut.edu.cn*

LWDNT-VGGM exploits lightweight deep networks for visual tracking. A lightweight fully convolutional network based on VGG-M-2048 is designed and trained on the ILSVRC VID dataset using mutual learning (between VGG-M and VGG-16). In online tracking, the proposed model outputs a response map regarding the target, based on which the target can be located by finding the peak of the response map. Besides, the scale estimation scheme proposed in DSST [12] is used.

### A.42 Lightweight Deep Neural Network for Visual Tracking (LWDNTthi)

*H. Zhao, D. Wang, H. Lu*
*zhaohj@stumail.neu.edu.cn, {wdice, lhchuan}@dlut.edu.cn*

LWDNTthi exploits lightweight deep networks for visual tracking. To be specific, a lightweight fully convolutional network based on ThiNet is designed, and it is trained on the ILSVRC VID dataset directly. In online tracking, our model outputs a response map regarding the target, based on which the target can be located by finding the peak of the response map. The scale estimation scheme proposed in DSST [12] is also used.

### A.43 Local-Global Tracking tracker (LGT)

*Submitted by VOT Committee*

The core element of LGT is a coupled-layer visual model that combines the target global and local appearance by interlacing two layers. By this coupled constraint paradigm between the adaptation of the global and the local layer, a more robust tracking through significant appearance changes is achieved. The reader is referred to [82] for details.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### A.44 L1APG (L1APG)

*Submitted by VOT Committee*

L1APG [4] considers tracking as a sparse approximation problem in a particle filter framework. To find the target in a new frame, each target candidate is sparsely represented in the space spanned by target templates and trivial templates. The candidate with the smallest projection error after solving an $\ell_1$ regularized least squares problem. The Bayesian state inference framework is used to propagate sample distributions over time.

### A.45 Matrioska (Matrioska)

*M. E. Maresca, A. Petrosino*
*mariomaresca@hotmail.it, alfredo.petrosino@uniparthenope.it*

The Matrioska's confidence score is based on the number of keypoints found inside the object in the initialization.

### A.46 Matrioska Best Displacement Flow (Matflow)

*M. E. Maresca, A. Petrosino*
*mariomaresca@hotmail.it, alfredo.petrosino@uniparthenope.it*

MatFlow enhances the performance of the first version of Matrioska [59] with response given by the short-term tracker BDF (see A.5).

### A.47 MEEM (MEEM)

*Submitted by VOT Committee*

MEEM [97] uses an online SVM with a re-detection based on the entropy of the score function. The tracker creates an ensamble of experts by storing historical snapshots while tracking. When needed the tracker can be restored by the best of these experts, selected using an entropy minimization criterion.

### A.48 MobileNet combined with SiameseFC (MBSiam)

*Y. Zhang, L. Wang, D. Wang, H. Lu*
*{zhangyunhua, wlj}@mail.dlut.edu.cn, {wdice, lhchuan}@dlut.edu.cn*

MBSiam uses a bounding box regression network to assist SiameseFC during online tracking. SiameseFC determines the center of the target and the size of the target is further predicted by the bounding box regression network. The SiameseFC network is similar to Bertinetto's work [6] using AlexNet architecture. Bounding box regression network uses SSD-MobileNet architecture [35,52] and it aims to regress the tight bounding box of the target object in a region during tracking given the target's appearance in the first frame.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### A.49 Multi Rotate and Scale Normalized Cross Correlation tracker (MRSNCC)

*A. G. Perera, Y. W. Law, J. Chahl*
*asanka.perera@mymail.unisa.edu.au, {yeewei.law, javaan.chahl}@unisa.edu.au*

The tracker MRSNCC performs multiple stages of rotation and scaling up and down to the region of interest. The target location is localized with a normalized cross correlation filter. This tracking is combined with a corner point tracker and a histogram based tracker to handle low confident estimations.

### A.50 Multi-Cue Correlation Tracker (MCCT)

*N. Wang, W. Zhou, H. Li*
*wn6149@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn*

The multi-cue correlation tracker (MCCT) is based on the discriminative correlation filter framework. By combining different types of features, the proposed approach constructs multiple experts and each of them tracks the target independently. With the proposed robustness evaluation strategy, the suitable expert is selected for tracking in each frame. Furthermore, the divergence of multiple experts reveals the reliability of the current tracking, which helps updating the experts adaptively to keep them from corruption.

### A.51 Multi-solution Fusion for Visual Tracking (MFT)

*S. Bai, Z. He, J. Zhuang*
*{baishuai, he010103}@bupt.edu.cn, junfei.zhuang@faceall.cn*

MFT tracker is based on correlation filtering algorithm. Firstly, different multi-resolution features with continuous convolution operator [15] are combined. Secondly, in order to improve the robustness a multi-solution using different features is trained and multi-solutions are optimally fused to predict the target location. Lastly, different combinations of Res50, SE-Res50, Hog, and CN features are applied to the different tracking situations.

### A.52 Multiple Instance Learning tracker (MIL)

*Submitted by VOT Committee*

MIL tracker [3] uses a tracking-by-detection approach, more specifically Multiple Instance Learning instead of traditional supervised learning methods and shows improved robustness to inaccuracies of the tracker and to incorrectly labelled training samples.

### A.53 Online Adaptive Hidden Markov Model for Multi-Tracker Fusion (HMMTxD)

*Submitted by VOT Committee*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



The HMMTxD method fuses observations from complementary out-of-the box trackers and a detector by utilizing a hidden Markov model whose latent states correspond to a binary vector expressing the failure of individual trackers. The Markov model is trained in an unsupervised way, relying on an online learned detector to provide a source of tracker-independent information for a modified Baum-Welch algorithm that updates the model w.r.t. the partially annotated data.

### A.54 Part-based tracking by sampling (PBTS)

*George De Ath, Richard Everson*
{*gd295, r.m.everson*}*@exeter.ac.uk*

PBTS [18] describes objects with a set of image patches which are represented by pairs of RGB pixel samples and counts of how many pixels in the patch are similar to them. This empirically characterises the underlying colour distribution of the patches and allows for matching using the Bhattacharyya distance. Candidate patch locations are generated by applying non-shearing affine transforms to the patches' previous locations, which are then evaluated for their match quality, and the best of these are locally optimised in a small region around each patch.

### A.55 Robust Fragments based Tracking using the Integral Histogram - FragTrack (FT)

*Submitted by VOT Committee*

FragTrack represents the model of the object by multiple image fragments or patches. The patches are arbitrary and are not based on an object model. Every patch votes on the possible positions and scales of the object in the current frame, by comparing its histogram with the corresponding image patch histogram. A robust statistic is minimized in order to combine the vote maps of the multiple patches. The algorithm overcomes several difficulties which cannot be handled by traditional histogram-based algorithms like partial occlusions or pose change.

### A.56 Robust Multi-task Correlation Particle Filter (R_MCPF)

*J. Gao, T. Zhang, Y. Jiao, C. Xu*
{*gaojunyu2012, yifanjiao1227*}*@gmail.com,* {*tzzhang, csxu*}*@nlpr.ia.ac.cn*

R_MCPF is based on the MCPF tracker [98] with a more robust fusion strategy for deep features.

### A.57 ROI-Align Network (RAnet)

*S. Yun, D. Wee, M. Kang, J. Sung*
{*sangdoo.yun, dongyoon.wee, myunggu.kang, jinyoung.sung*}*@navercorp.com*

This tracker is based on tracking-by-detection approach using CNNs. To make the tracker faster, a new tracking framework using RoIAlign technique is proposed.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### A.58 Salient Region weighted Correlation filter Tracker (SRCT)

*H. Lee, D. Kim*
{*lhmin, dkim*}*@postech.ac.kr*

SRCT is the ensemble tracker composed of Salient Region-based Tracker [46] and ECO tracker [15]. The score map of Salient Region based Tracker is weighted to the score map of ECO tracker in spatial domain.

### A.59 SA_Siam_P - An Advanced Twofold Siamese Network for Real-Time Object Tracking (SA_Siam_P)

*A. He, C. Luo, X. Tian, W. Zeng*
*heanfeng@mail.ustc.edu.cn,* {*cluo, wezeng*}*@microsoft.com, xinmei@ustc.edu.cn*

SA_Siam_P is an implementation of the SA-Siam tracker as described in [30]. Some bugs in the original implementation were fixed. In addition, for sequences where the target bounding box is not upright in the first frame, the reported tracking results are bounding boxes with the same tilt angle as the box in the first frame.

### A.60 SA_Siam_R: A Twofold Siamese Network for Real-Time Object Tracking With Angle Estimation (SA_Siam_R)

*A. He, C. Luo, X. Tian, W. Zeng*
*heanfeng@mail.ustc.edu.cn,* {*cluo, wezeng*}*@microsoft.com, xinmei@ustc.edu.cn*

SA_Siam_R is a variation of the Siamese network-based tracker SA-Siam [30]. SA_Siam_R adopts three simple yet effective mechanisms, namely angle estimation, spatial mask, and template update, to achieve a better performance than SA-Siam. First, the framework includes multi-scale multi-angle candidates for search region. The scale change and the angle change of the tracked object are implicitly estimated according to the response maps. Second, spatial mask is applied when the aspect ratio of the target is apart from 1:1 to reduce background noise. Last, moving average template update is adopted to deal with hard sequences with large target deformation. For more details, the reader is referred to [29].

### A.61 Scale Adaptive Mean-Shift Tracker (ASMS)

*Submitted by VOT Committee*

The mean-shift tracker optimizes the Hellinger distance between template histogram and target candidate in the image. This optimization is done by a gradient descend. ASMS [87] addresses the problem of scale adaptation and presents a novel theoretically justified scale estimation mechanism which relies solely on the mean-shift procedure for the Hellinger distance. ASMS also introduces two improvements of the mean-shift tracker that make the scale estimation more robust in the presence of background clutter – a novel histogram colour weighting and a forward-backward consistency check. Code available at https://github.com/vojirt/asms.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### A.62 Scale Adaptive Point-based Kanade Lukas Tomasi colour-Filter (SAPKLTF)

*R. Martín-Nieto, Á. García-Martín, J. M. Martínez, Á. Iglesias-Arias, P. Vicente-Moñivar, S. Vivas, E. Velasco-Salido*
{*rafael.martinn, alvaro.garcia, josem.martinez, alvaro.iglesias, pablo.vicente, sergio.vivas, erik.velasco*}*@uam.es*

The SAPKLTF [85] tracker is based on an extension of PKLTF tracker [24] with ASMS [87]. SAPKLTF is a single-object long-term tracker which consists of two phases: The first stage is based on the Kanade Lukas Tomasi approach (KLT) [73] choosing the object features (colour and motion coherence) to track relatively large object displacements. The second stage is based on scale adaptive mean shift gradient descent [87] to place the bounding box into the exact position of the object. The object model consists of a histogram including the quantized values of the RGB colour components and an edge binary flag.

### A.63 SiamVGG (SiamVGG)

*Y. Li, C. Hao, X. Zhang, H. Zhang, D. Chen*
*leeyh@illinois.edu, hc.onioncc@gmail.com, xiaofan3@illinois.edu, zhhg@bupt.edu.cn, dchen@illinois.edu*

SiamVGG adopts SiamFC [6] as the baseline approach. It applies a fully-convolutional Siamese network to allocate the target in the search region using a modified VGG-16 network [74] as the backbone. The network is trained offline on both ILSVRC VID dataset [71] and Youtube-BB dataset end-to-end.

### A.64 Spatially Regularized Discriminative Correlation Filter Tracker (SRDCF)

*Submitted by VOT Committee*

Standard Discriminative Correlation Filter (DCF) based trackers such as [12,32,14] suffer from the inherent periodic assumption when using circular correlation. The Spatially Regularized DCF (SRDCF) alleviates this problem by introducing a spatial regularization function that penalizes filter coefficients residing outside the target region. This allows the size of the training and detection samples to be increased without affecting the effective filter size. By selecting the spatial regularization function to have a sparse Discrete Fourier Spectrum, the filter is efficiently optimized directly in the Fourier domain. For more details, the reader is referred to [13].

### A.65 Spatio-Temporal Background-Aware Correlation Filter for Visual Tracking (STBACF)

*A. Memarmoghdam, H. Kiani Galoogah*
*a.memarmoghadam@eng.ui.ac.ir, hamedkg@gmail.com*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



Recently, the discriminative BACF approach [23] efficiently tracks the target object via training a correlation filter by exploiting real negative examples densely sampled from its surrounding background. To further improve its robustness, especially against drastic changes of the object model during track, STBACF tracker simultaneously updates the filter while training by incorporating temporal regularization into the original BACF formulation. In this way, a temporally consistent filter is efficiently solved in each frame via an iterative ADMM method. Furthermore, to suppress unwanted non-object information of the target bounding box, an elliptical binary mask is applied during online training.

### A.66 Spatio-temporal Siamese Tracking (STST)

*F. Zhao, Y. Wu, J. Wang, M. Tang*
{*fei.zhao, jqwang, tangm*}*@nlpr.ia.ac.cn, ywu.china@gmail.com*

The tracker STST applies 3D convolutional block to extract the temporal features of the target appearing in different frames, and it uses the dense correlation layer to match the feature maps of the target patch and the search patch.

### A.67 Staple: Sum of Template And Pixel-wise LEarners (Staple)

*L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P. Torr*
{*luca.bertinetto, stuart.golodetz, ondrej.miksik, philip.torr*}*@eng.ox.ac.uk,*
*jack.valmadre@gmail.com*

Staple is a tracker that combines two image patch representations that are sensitive to complementary factors to learn a model online that is inherently robust to both colour changes and deformations. For more details, we refer the reader to [7].

### A.68 Struck: Structured output tracking with kernels (struck2011)

*Submitted by VOT Committee*

Struck [28] is a framework for adaptive visual object tracking based on structured output prediction. The method uses a kernelized structured output support vector machine (SVM), which is learned online to provide adaptive tracking.

### A.69 TRAcker based on Context-aware deep feature compression with multiple Auto-encoders (TRACA)

*J. Choi, H. J. Chang, T. Fischer, S. Yun, Y. Demiris, J. Y. Choi*
*jwchoi.pil@gmail.com,* {*hj.chang, t.fischer, y.demiris*}*@imperial.ac.uk,*
{*yunsd101, jychoi*}*@snu.ac.kr*

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



The proposed TRACA consists of multiple expert auto-encoders, a context-aware network, and correlation filters. The expert auto-encoders robustly compress raw deep convolutional features from VGG-Net. Each of them is trained according to a different context, and thus performs context-dependent compression. A context-aware network is proposed to select the expert auto-encoder best suited for the specific tracking target. During online tracking, only this auto-encoder is running. After initially adapting the selected expert auto-encoder for the tracking target, its compressed feature map is utilized as an input of correlation filters which tracks the target online.

### A.70 Tracking by Feature Select Adversary Network (FSAN)

*W. Wei, Q. Ruihe, L. Si*
*wang_wei.buaa@163.com, {qianruihe, liusi}@iie.ac.cn*

The tracker FSAN consists of an offline trained convolutional network and a feature channels selecting adversary network. Image patches are extracted and multiple channels feature of each patch in each frame are computed. Then, the more stable discriminative feature in is selected by a channel mask generate network. The generate network can filter out the most discriminative feature channels in current frame. In the adversarial learning, the robustness of the discriminative network is increased by using examples in which the feature channels are enhanced or removed by the generate network.

### A.71 Unveiling the Power of Deep Tracking (UPDT)

*G. Bhat, J. Johnander, M. Danelljan, F. Khan, M. Felsberg*
*{goutam.bhat, joakim.johnander, martin.danelljan, fahad.khan, michael.felsberg}@liu.se*

UPDT learns independent tracking models for deep and shallow features to fully exploit their complementary properties. The deep model is trained with an emphasis on achieving higher robustness, while the shallow model is trained to achieve high accuracy. The scores of these individual models are then fused using a maximum margin based approach to get the final target prediction. For more details, the reader is referred to [8].

### A.72 3D Convolutional Networks for Visual Tracking (C3DT)

*H. Li, S. Wu, Y. Yang, S. Huang*
*haojieli_scut@foxmail.com, eesihang@mail.scut.edu.cn, yychzw@foxmail.com, eehsp@scut.edu.cn*

The tracker C3DT improves the existing tracker MDNet [64] by introducing spatio-temporal information using the C3D network [80]. MDNet treats the tracking as classification and regression, which utilizes the appearance feature from the current frame to determine which candidate frame is object or background, and then gets an accurate bounding box by a linear regression. This

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



network ignores the importance of spatio-temporal information for visual tracking. To address this problem C3DT tracker adopts two-branch network to extract features. One branch is used to get features from the current frame by the VGG-S [11]; another is the C3D network, which extracts spatio-temporal information from the previous frames.

## B VOT2018 long-term challenge tracker descriptions

In this appendix we provide a short summary of all trackers that were submitted to the long-term challenge.

### B.1 A Memory Model based on the Siamese Network for Long-term Tracking (MMLT)

*H. Lee, S. Choi, C. Kim*
{*hankyeol, seokeon, changick*}*@kaist.ac.kr*

MMLT consists of three parts: memory management, tracking, and re-detection. The structure of the memory model for long-term tracking, which is inspired by the well-known Atkinson-Shiffrin model [2], is divided into the short-term and long-term stores. Tracking and re-detection processes are performed based on this memory model. In the tracking step, the bounding box of the target is estimated by combining the features of the Siamese network [6] in both short-term and long-term stores. In the re-detection step, features in the long-term store are employed. A coarse-to-fine strategy is adopted that collects candidates with similar semantic meanings in the entire image and then it refines the final position based on the Siamese network.

### B.2 DaSiameseRPN_long-term (DaSiam_LT)

*Z. Zhu, Q. Wang, B. Li, W. Wu, Wei Zou*
{*zhuzheng2014, wangqiang2015, wei.zou*}*@ia.ac.cn,* {*libo, wuwei*}*@sensetime.com*

The tracker DaSiam_LT adopts Siamese Region Proposal Network (SiamRPN) A.35 as the baseline. It extends the SiamRPN approach by introducing a simple yet effective local-to-global search region strategy. Specifically, the size of search region is iteratively growing with a constant step when failed tracking is indicated. The distractor-aware training and inference are added to enable high-quality detection score to indicate the quality of tracking results [102].

### B.3 Flock of Trackers (FoT)

*Submitted by VOT Committee*

For a tracker description, the reader is referred to A.33.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### B.4 Fully-Convolutional Siamese Detector (SiamFCDet)

*J. Valmadre, L. Bertinetto, N. Lee, J. Henriques, A. Vedaldi, P. Torr*
*jack.valmadre@gmail.com, {luca.bertinetto, namhoon.lee, joao.henriques, andrea.vedaldi, philip.torr}@eng.ox.ac.uk*

SiamFCDet uses SiamFC to search the entire image at multiple resolutions in each frame. There is no temporal component.

### B.5 Fully-Convolutional Siamese Network (SiamFC)

*J. Valmadre, L. Bertinetto, N. Lee, J. Henriques, A. Vedaldi, P. Torr*
*jack.valmadre@gmail.com, {luca.bertinetto, namhoon.lee, joao.henriques, andrea.vedaldi, philip.torr}@eng.ox.ac.uk*

For a tracker description, the reader is referred to A.34.

### B.6 Fully Correlational Long-Term Tracker (FuCoLoT)

*Submitted by VOT Committee*

FuCoLoT is a Fully Correlational Long-term Tracker. It exploits the novel DCF constrained filter learning method to design a detector that is able to re-detect the target in the whole image efficiently. Several correlation filters are trained on different time scales that act as the detector components. A mechanism based on the correlation response is used for tracking failure estimation.

### B.7 Long-Term Siamese Instance Search Tracking (LTSINT)

*R. Tao, E. Gavves, A. Smeulders*
*{rantao.mail, efstratios.gavves}@gmail.com, a.w.m.smeulders@uva.nl*

The tracker follows the Siamese tracking framework. It has two novel components. One is a hybrid search scheme which combines local search and global search. The global search is a three-step procedure following a coarse-to-fine scheme. The tracker switches from local search to global search when the similarity score of the detected box is below a certain threshold (0.3 for this submission). The other novel component is a cautious model updating which updates the similarity function online. Model updates are permissible when the similarity score of the detected box is above a certain threshold (0.5 for this submission).

### B.8 MobileNet based tracking by detection algorithm (MBMD)

*Y. Zhang, L. Wang, D. Wang, J. Qi, H. Lu*
*{zhangyunhua, wlj}@mail.dlut.edu.cn, {wdice, jinqing, lhchuan}@dlut.edu.cn*

The proposed tracker consists of a bounding box regression network and a verifier network. The regression network regresses the target object's bounding box in a search region given the target in the first frame. Its outputs are several candidate boxes and each box's reliability is evaluated by the verifier to

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



determine the predicted target box. If the predicted scores of both networks are below the thresholds, the tracker searches the target in the whole image. The regression network uses SSD-MobileNet architecture [35,52] and its parameters are fixed during online tracking. The verifier is similar to MDNet [64] and is implemented by VGGM pretrained on ImageNet classification dataset. The last three layers' parameters of the verifier are updated online to filter the distractors for the tracker.

### B.9 Online Adaptive Hidden Markov Model for Multi-Tracker Fusion (HMMTxD)

*Submitted by VOT Committee*

For a tracker description, the reader is referred to A.53.

### B.10 Parallel Tracking and Verifying Plus (PTAVplus)

*H. Fan, F. Yang, Q. Zhou, H. Ling*
{*hengfan, fyang, hbling*}*@temple.edu, zhou.qin.190@sjtu.edu.cn*

PTAVplus is an improvement of PTAV [20] by combining a tracker and a strong verifier for long-term visual tracking.

### B.11 Scale Adaptive Mean-Shift Tracker (ASMS)

*Submitted by VOT Committee*

For a tracker description, the reader is referred to A.61.

### B.12 Scale Adaptive Point-based Kanade Lukas Tomasi colour-Filter (SAPKLTF)

*R. Martín-Nieto, Á. García-Martín, J. M. Martínez, Á. Iglesias-Arias, P. Vicente-Moñivar, S. Vivas, E. Velasco-Salido*
{*rafael.martinn, alvaro.garcia, josem.martinez, alvaro.iglesias, pablo.vicente, sergio.vivas, erik.velasco*}*@uam.es*

For a tracker description, the reader is referred to A.62.

### B.13 Search your object with siamese network (SYT)

*P. Li, Z. Wang, D. Wang, B. Chen, H. Lu*
{*907508458, 2805825263*}*@qq.com, wdice@dlut.edu.cn, 476732833@qq.com, lhchuan@dlut.edu.cn*

In long-term tracking, few trackers can re-detect the object after tracking failures. SYT utilises the siamese network as base tracker and it introduces the Single Shot MultiBox Detector for re-detection. A verifier with the initial frame to output the tracking score is trained. When the score is larger than zero, the tracker result is utilised; otherwise, the detector to re-find the object in the whole images is used.

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



### B.14 Siamese Long-term Tracker (SLT)

*J. Zhuang, S. Bai, Z. He*
*junfei.zhuang@facell.cn, {baishuai, he010103}@bupt.edu.cn*

Siamese Long-term tracker (SLT) is composed of two main components. The first part is short-term tracker based on SiamFC-3s [6]. The role of this part is tracking target before it disappears from view. The second part is a detector which aims to re-detect the target when it reappears, and it is also based on Siamese network structure. For this part, a modified VGG-M model is employed to extract target features from the first frame and whole image features from other frames, then target features are compared with whole image features to locate target position in a new frame.

### B.15 SiamVGG (SiamVGG)

*Y. Li, C. Hao, X. Zhang, H. Zhang, D. Chen*
*leeyh@illinois.edu, hc.onioncc@gmail.com, xiaofan3@illinois.edu, zhhg@bupt.edu.cn, dchen@illinois.edu*

For a tracker description, the reader is referred to A.63.

## References

1. Abdelpakey, M.H., Shehata, M.S., Mohamed, M.M.: Denssiam: End-to-end densely-siamese network with self-attention model for object tracking. ArXiv e-prints (Sep 2018)
2. Atkinson, R.C., Shiffrin, R.M.: Human memory: A proposed system and its control processes1. In: Psychology of learning and motivation, vol. 2, pp. 89–195. Elsevier (1968)
3. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1619–1632 (2011)
4. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust l1 tracker using accelerated proximal gradient approach. In: CVPR (2012)
5. Battistone, F., Petrosino, A., Santopietro, V.: Watch out: Embedded video tracking with BST for unmanned aerial vehicles. Journal of Signal Processing Systems **90**(6), 891–900 (Jun 2018). https://doi.org/10.1007/s11265-017-1279-x, https://doi.org/10.1007/s11265-017-1279-x
6. Bertinetto, L., Valmadre, J., Henriques, J., Torr, P.H.S., Vedaldi, A.: Fully convolutional siamese networks for object tracking. In: ECCV Workshops. pp. 850–865 (2016)
7. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S.: Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1401–1409 (2016)
8. Bhat, G., Johnander, J., Danelljan, M., Khan, F.S., Felsberg, M.: Unveiling the power of deep tracking. In: ECCV (2018)
9. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



10. Čehovin, L.: TraX: The visual Tracking eXchange Protocol and Library. Neurocomputing (2017). https://doi.org/http://dx.doi.org/10.1016/j.neucom.2017.02.036
11. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman., A.: Return of the devil in the details: Delving deep into convolutional nets. In: BMVC (2014)
12. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference BMVC (2014)
13. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: International Conference on Computer Vision (2015)
14. Danelljan, M., Khan, F.S., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: Computer Vision and Pattern Recognition (2014)
15. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ECO: efficient convolution operators for tracking. In: CVPR (2017)
16. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2016)
17. Danelljan, M., Robinson, A., Shahbaz Khan, F., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: ECCV. pp. 472–488 (2016)
18. De Ath, G., Everson, R.: Part-Based Tracking by Sampling. ArXiv e-prints (May 2018)
19. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR (2009)
20. Fan, H., Ling, H.: Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In: ICCV (2017)
21. Felsberg, M., Berg, A., Häger, G., Ahlberg, J., et al.: The thermal infrared visual object tracking VOT-TIR2015 challenge results. In: ICCV2015 workshop proceedings, VOT2015 Workshop (2015)
22. Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. CoRR **abs/1703.05884** (2017), http://arxiv.org/abs/1703.05884
23. Galoogahi, H.K., Fagg, A., Lucey, S.: Learning background-aware correlation filters for visual tracking. In: ICCV. pp. 1144–1152 (2017)
24. González, A., Martín-Nieto, R., Bescós, J., Martínez, J.M.: Single object long-term tracker for smart control of a ptz camera. In: Proceedings of the International Conference on Distributed Smart Cameras. p. 39. ACM (2014)
25. Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: A new change detection benchmark dataset. In: CVPR Workshops. pp. 1–8. IEEE (2012)
26. Gundogdu, E., Alatan, A.A.: Good features to correlate for visual tracking. IEEE Transactions on Image Processing **27**(5), 2526–2540 (May 2018). https://doi.org/10.1109/TIP.2018.2806280
27. Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S.: Learning dynamic Siamese network for visual object tracking. In: ICCV (2017)
28. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V. (eds.) International Conference on Computer Vision. pp. 263–270. IEEE (2011)

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



29. He, A., Luo, C., Tian, X., Zeng, W.: Towards a better match in siamese network based visual object tracker. In: The Visual Object Tracking (VOT) Challenge Workshop in the European Conference on Computer Vision (ECCV) (September 2018)
30. He, A., Luo, C., Tian, X., Zeng, W.: A twofold siamese network for real-time object tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
31. He, Z., Fan, Y., Zhuang, J., Dong, Y., Bai, H.: Correlation filters with weighted convolution responses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1992–2000 (2017)
32. Henriques, J., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. PAMI **37**(3), 583–596 (2015)
33. Herranz, J.R.: Short-term single target tracking with discriminative correlation filters. Master thesis, University of Ljubljana/Technical University of Madrid (2018)
34. Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D.: Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 749–758 (2015)
35. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. In: arXiv preprint arXiv:1704.04861 (2017)
36. Jack, V., Luca, B., ao F., H.J., Ran, T., Andrea, V., Arnold, S., Philip, T., Efstratios, G.: Long-term tracking in the wild: A benchmark. arXiv:1803.09502 (2018)
37. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(7), 1409–1422 (2012). https://doi.org/10.1109/TPAMI.2011.239
38. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojíř, T., Häger, G., Lukežič, A., Eldesokey, A., Fernández, G., et al.: The visual object tracking vot2017 challenge results. In: ICCV2017 Workshops, Workshop on visual object tracking challenge (2017)
39. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojíř, T., Häger, G., Lukežič, A., Fernández, G., et al.: The visual object tracking vot2016 challenge results. In: ECCV2016 Workshops, Workshop on visual object tracking challenge (2016)
40. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Čehovin, L., Fernández, G., Vojíř, T., Häger, G., Nebehay, G., Pflugfelder, R., et al.: The visual object tracking vot2015 challenge results. In: ICCV2015 Workshops, Workshop on visual object tracking challenge (2015)
41. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Čehovin, L., Nebehay, G., G., F., Vojíř, T., et al.: The visual object tracking vot2013 challenge results. In: ICCV2013 Workshops, Workshop on visual object tracking challenge. pp. 98 –111 (2013)
42. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Čehovin, L., Nebehay, G., Vojíř, T., Fernández, G., et al.: The visual object tracking vot2014 challenge results. In: ECCV2014 Workshops, Workshop on visual object tracking challenge (2014)
43. Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernández, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2137–2155 (2016)

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



44. Leal-Taixé, L., Milan, A., Reid, I.D., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. CoRR **abs/1504.01942** (2015), http://arxiv.org/abs/1504.01942
45. Lebeda, K., Bowden, R., Matas, J.: Long-term tracking through failure cases. In: Vis. Obj. Track. Challenge VOT2013, In conjunction with ICCV2013 (2013)
46. Lee, H., Kim, D.: Salient region-based online object tracking. In: Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on. pp. 1170–1177. IEEE (2018)
47. Li, A., Li, M., Wu, Y., Yang, M.H., Yan, S.: Nus-pro: A new visual tracking challenge. IEEE-PAMI (2015)
48. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
49. Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.H.: Learning spatial-temporal regularized correlation filters for visual tracking. In: CVPR (2018)
50. Li, Y., Zhu, J., Song, W., Wang, Z., Liu, H., Hoi, S.C.H.: Robust estimation of similarity transformation for visual object tracking with correlation filters (2017)
51. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE Transactions on Image Processing **24**(12), 5630–5644 (2015)
52. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
53. Lukežič, A., Vojíř, T., Zajc, L.Č., Matas, J., Kristan, M.: Discriminative correlation filter with channel and spatial reliability. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6309–6318 (July 2017)
54. Lukežič, A., Č. Zajc, L., Kristan, M.: Deformable parts correlation filters for robust visual tracking. IEEE Transactions on Cybernetics **PP**(99), 1–13 (2017)
55. Lukezic, A., Zajc, L.C., Vojír, T., Matas, J., Kristan, M.: FCLT - A fully-correlational long-term tracker. CoRR **abs/1711.09594** (2017), http://arxiv.org/abs/1711.09594
56. Lukezic, A., Zajc, L.C., Vojír, T., Matas, J., Kristan, M.: Now you see me: evaluating performance in long-term visual tracking. CoRR **abs/1804.07056** (2018), http://arxiv.org/abs/1804.07056
57. Ma, C., Yang, X., Zhang, C., Yang, M.H.: Long-term correlation tracking. In: CVPR (2015)
58. Maresca, M., Petrosino, A.: Clustering local motion estimates for robust and efficient object tracking. In: Proceedings of the Workshop on Visual Object Tracking Challenge, European Conference on Computer Vision (2014)
59. Maresca, M.E., Petrosino, A.: Matrioska: A multi-level approach to fast tracking by learning. In: Proc. Int. Conf. Image Analysis and Processing. pp. 419–428 (2013)
60. Moudgil, A., Gandhi, V.: Long-term visual object tracking benchmark. arXiv preprint arXiv:1712.01358 (2017)
61. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: European Conference on Computer Vision. pp. 445–461 (2016)
62. Müller, M., Bibi, A., Giancola, S., Al-Subaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. CoRR **abs/1803.10794** (2018), http://arxiv.org/abs/1803.10794
63. N. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 1, pp. 886–893 (June 2005)

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



64. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: CVPR. pp. 4293–4302 (2016)
65. Nebehay, G., Pflugfelder, R.: Clustering of Static-Adaptive correspondences for deformable object tracking. In: Computer Vision and Pattern Recognition. IEEE (2015)
66. Pernici, F., del Bimbo, A.: Object tracking by oversampling local features. IEEE Trans. Pattern Anal. Mach. Intell. **36**(12), 2538–2551 (2013). https://doi.org/http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.250
67. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
68. Possegger, H., Mauthner, T., Bischof, H.: In defense of color-based model-free tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
69. Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V.: YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video. In: Computer Vision and Pattern Recognition. pp. 7464–7473 (2017)
70. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision **77**(1-3), 125–141 (2008)
71. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y, http://dx.doi.org/10.1007/s11263-015-0816-y
72. Senna, P., Drummond, I.N., Bastos, G.S.: Real-time ensemble-based tracker with kalman filter. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). pp. 338–344 (Oct 2017). https://doi.org/10.1109/SIBGRAPI.2017.51
73. Shi, J., Tomasi, C.: Good features to track. In: Computer Vision and Pattern Recognition. pp. 593 – 600 (June 1994)
74. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
75. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual Tracking: an Experimental Survey. TPAMI (2013). https://doi.org/10.1109/TPAMI.2013.230
76. Solera, F., Calderara, S., Cucchiara, R.: Towards the evaluation of reproducible robustness in tracking-by-detection. In: Advanced Video and Signal Based Surveillance. pp. 1 – 6 (2015)
77. Sun, C., Wang, D., Lu, H., Yang, M.H.: Correlation tracking via joint discrimination and reliability learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 489–497 (2018)
78. Sun, C., Wang, D., Lu, H., Yang, M.H.: Learning spatial-aware regressions for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8962–8970 (2018)
79. Tao, R., Gavves, E., Smeulders, A.W.M.: Tracking for half an hour. CoRR **abs/1711.10217** (2017), http://arxiv.org/abs/1711.10217
80. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497 (Dec 2015). https://doi.org/10.1109/ICCV.2015.510

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



81. Valmadre, J., Bertinetto, L., Henriques, J.F., Vedaldi, A., Torr, P.H.: End-to-end representation learning for correlation filter based tracking. arXiv preprint arXiv:1704.06036 (2017)
82. Čehovin, L., Kristan, M., Leonardis, A.: Robust visual tracking using an adaptive coupled-layer visual model. IEEE Trans. Pattern Anal. Mach. Intell. **35**(4), 941–953 (2013). https://doi.org/http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.145
83. Čehovin, L., Leonardis, A., Kristan, M.: Visual object tracking performance measures revisited. IEEE Transactions on Image Processing **25**(3) (2015)
84. Čehovin, L., Leonardis, A., Kristan, M.: Robust visual tracking using template anchors. In: WACV. IEEE (Mar 2016)
85. Velasco-Salido, E., Martínez, J.M.: Scale adaptive point-based kanade lukas tomasi colour-filter tracker. Under Review (2017)
86. Vojíř, T., Matas, J.: The enhanced flock of trackers. In: Cipolla, R., Battiato, S., Farinella, G.M. (eds.) Registration and Recognition in Images and Videos, Studies in Computational Intelligence, vol. 532, pp. 113–136. Springer Berlin Heidelberg, Springer Berlin Heidelberg (January 2014). https://doi.org/10.1007/978-3-642-44907-9_6, http://dx.doi.org/10.1007/978-3-642-44907-9_6
87. Vojíř, T., Noskova, J., Matas, J.: Robust scale-adaptive mean-shift for tracking. Pattern Recognition Letters **49**, 250–258 (2014)
88. Wang, Q., Gao, J., Xing, J., Zhang, M., Hu, W.: Dcfnet: Discriminant correlation filters network for visual tracking. arXiv preprint arXiv:1704.04057 (2017)
89. Van de Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing **18**(7), 1512–1523 (2009)
90. Wu, C., Zhu, J., Zhang, J., Chen, C., Cai, D.: A convolutional treelets binary feature approach to fast keypoint recognition. In: ECCV. pp. 368–382 (2013)
91. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Computer Vision and Pattern Recognition (2013)
92. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. PAMI **37**(9), 1834–1848 (2015)
93. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking. arXiv preprint arXiv:1807.11348 (2018)
94. Yiming, L., Shen, J., Pantic, M.: Mobile face tracking: A survey and benchmark. arXiv:1805.09749v1 (2018)
95. Young, D.P., Ferryman, J.M.: PETS Metrics: On-line performance evaluation service. In: ICCCN '05 Proceedings of the 14th International Conference on Computer Communications and Networks. pp. 317–324 (2005)
96. Zajc, L.Č., Lukežič, A., Leonardis, A., Kristan, M.: Beyond standard benchmarks: Parameterizing performance evaluation in visual object tracking. ICCV **abs/1612.00089** (2017), http://arxiv.org/abs/1612.00089
97. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimization. In: ECCV (2014)
98. Zhang, T., Xu, C., Yang, M.H.: Learning multi-task correlation particle filters for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–14 (2018)
99. Zhang, Z., Li, Y., Ren, J., Zhu, J.: Effective occlusion handling for fast correlation filter-based trackers (2018)
100. Zhu, G., Porikli, F., Li, H.: Tracking randomly moving objects on edge box proposals. CoRR (2015)

European Conference on Computer Vision Workshops (2018), pp: 3-53
Preprint version; final version available at DOI: 10.1007/978-3-030-11009-3_1



101. Zhu, P., Wen, L., Bian, X., Haibin, L., Hu, Q.: Vision meets drones: A challenge. arXiv preprint arXiv:1804.07437 (2018)
102. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision (2018)